VISTAS IN ASTRONOMY

Volume 4

VOL
4
196

VISTAS IN ASTRONOMY

EDITED BY

ARTHUR BEER

Volume 4

ASTRONOMICAL HISTORY
CELESTIAL MECHANICS
GEOPHYSICS
ARTIFICIAL SATELLITES
SOLAR RESEARCH
ASTROPHYSICS

PERGAMON PRESS
OXFORD · LONDON · NEW YORK · PARIS
1961

PERGAMON PRESS LTD.
Headington Hill Hall, Oxford
4 & 5 Fitzroy Square, London W.1

PERGAMON PRESS INC.
122 East 55th Street, New York 22, N.Y.
1404 New York Avenue N.W., Washington 5 D.C.

PERGAMON PRESS S.A.R.L.
*24 Rue des Écoles, Paris V^e^*

PERGAMON PRESS G.m.b.H.
Kaiserstrasse 75, Frankfurt am Main

First Published 1961

Copyright
©
1961
PERGAMON PRESS LTD.

Library of Congress Card No. 56–4213

Set in Modern No. 7 11 on 12 pt. and printed in Great Britain by
J. W. ARROWSMITH LTD., BRISTOL

INTRODUCTION

"... and finally I should like to express my respects to the shades of the scribes of Enūma-Anu-Enlil, descendents of Ekur-zākir or of Sin-leqē-unninni, and of all the other scribes who computed and wrote the texts which are published here. By their untiring efforts they built the foundations for the understanding of the laws of nature which our generation is applying so successfully to the destruction of civilization. Yet they also provided hours of peace for those who attempted to decode their lines of thought two thousand years later."

O. Neugebauer in "Astronomical Cuneiform Texts", vol. 1, Institute for Advanced Study, Princeton, N.J.; Lund Humphries, London 1955.

The Observatories,
Cambridge, 1961.

ARTHUR BEER.

VOL
4
196

CONTENTS OF VOLUME 4

page

An 8th-century Meridian Line: I-Hsing's Chain of Gnomons and the Pre-history of the Metric System *by* A. Beer. (The Observatories, Cambridge), Ho Ping-Yü (University of Malaya, Singapore), Lu Gwei-Djen (Newnham College, Cambridge), J. Needham, F.R.S. (Gonville and Caius College, Cambridge), E. G. Pulleyblank (Downing College, Cambridge), G. I. Thompson (Royal Observatory, Edinburgh) 3

Artificial Satellites *by* Sir Harrie S. W. Massey (University College, London) 29

Dynamical Effects in the Motion of Earth Sputniks *by* L. I. Sedov (Academician, Department of Physics, University of Moscow) 111

Photographic Observations of Artificial Earth Satellites *by* A. G. Massevitch and A. M. Lozinsky (The Astronomical Council of the USSR Academy of Sciences, Moscow) 118

Archaeomagnetism *by* J. C. Belshé (Department of Geodesy and Geophysics, University of Cambridge) 127

The Development of a Solar Centre of Activity *by* C. de Jager (Sterrewacht Sonnenborgh, Utrecht, Netherlands) 143

On the Dispersion in the Period-Luminosity Relation *by* C. Payne-Gaposchkin (Harvard College Observatory, Cambridge, Mass.) 184

Index 191

VOL
4
196

INDEX

Abéle, M. K., 123, 124, 126
Ainsworth, J., 66, 97, 99, 109
Aitken, M. J., 134, 141
Albus, J. S., 62, 109
Allen, C. W., 153, 179, 182, 183
Allen, M., 91
al-Ma'mūn, 4
Al'pert, Y. A., 102, 108
Archaeomagnetism, 127–142
Artificial satellites (see also Earth sputniks), 29–110, 88
Artificial satellites, air resistance, 38, 94
Artificial satellites, drag due to neutral molecules, 39
Artificial satellites, effect of the Earth's polar flattening, 50, 53, 99
Artificial satellites, electrical drag, 39
Artificial satellites, guidance conditions, 33
Artificial satellites, launching problems, 32
Artificial satellites, perturbations of orbits, 38, 41, 51, 53, 55, 56, 94
Artificial satellites, secular change of orbital inclination, 100
Artificial satellites, temperature of, 62
Āryabhaṭa, 19
Athay, R. G., 151, 182
Atmospheric pressure and density from observation of artificial satellites, 66, 97
Atmospheric pressure and density from observation of artificial satellites, electrical methods, 67
Atmospheric pressure and density from observation of artificial satellites, mechanical methods, 66

Babcock, H. D., 180, 181, 182
Babcock, H. W., 151, 168, 180, 181, 182, 183
Bain, W. C., 101, 108
Bartels, J., 74, 108, 128, 141, 181, 182
Bartman, F. L., 67, 108
Bates, D. R., 99, 108
Bauer, L. A., 128
Becker, U., 163, 182
Beer, A., 3, 23, 27, 109, 183
Beggs, D. W., 148, 182
Bellam, E. R., 101, 109
Belshé, J. C., 127, 141
Beynon, G., 182
Bidgood, D. E. T., 132, 141
Biermann, L., 147, 181, 182
Blitzer, L., 55, 108
Bogges, A., 109
Böhm, K. H., 148, 182
Boischot, A., 177, 178, 182
Booth, C. F., 101, 108
Borough, W., 127, 141
Bosmans, H., 27
Bracewell, R. N., 182, 183
Brown, G. M., 182
Bruzek, A., 149, 167, 182
Bullard, E. C., 128, 137, 141
Bumba, V., 168, 182
Butler, H. E., 92, 107
Byram, E. T., 109, 182

Cepheid variables, 184–189
Cepheid variables, period-frequencies, 187
Chapman, S., 74, 75, 108, 128, 141
Charitonov, A. V., 122, 126
Chêng Chung, 9
Chêng (Hsüan), 9
Chen Kuan, 10
Chevalier, R., 135, 141
Chhiu Chhang-Chhun, 14
Chhüthan Chuan, 25
Chhüthan (Gautama), 19
Chhüthan Hsi-Ta, 19, 25
Chiayeh (Kasyapa), 19
Chinese distance measurements, 16
Chinese geodetical surveys, 3–28
Chinese shadow observations, 17
Chingiz-Khan, 14
Chou-Kung, 8
Christofilos, N. C., 107, 108
Christiansen, W. N., 156, 182
Chubb, T. A., 69, 70, 108, 160, 178, 182
Chudakov, A. E., 105, 106, 108, 110
Chudesenko, E. F., 102, 108
Chu Kho-Chen, 14
Chümolo (Kumara), 19
Cook, R. M., 130, 141
Cooper, C., 142
Cornford, E. C., 56, 108
Corona and coronal activity centres, mass-balance of, 180
Cosmic and auroral radiations, studied with artificial satellites, 71
Cowling, T. G., 147, 152, 182
Crogg, Th., 168, 183

Danilin, B. S., 108
d'Anville, J. B. B., 16, 25, 27
Davis, R. J., 49, 108
de Jager, C., 143, 149, 163, 182
Denisse, J. F., 160, 177, 178, 182
Deutsch, A. N., 121, 126
Drummeter, L., 63, 108
Dubin, M., 108
Dzubenko, N. I., 180, 182

Earth's magnetic field, 127
Earth sputniks (see also artificial satellites), 29–110

Earth sputniks, dynamical effects in their motion, 111–117
Earth sputniks, photographic observations 118–126
Eckels, A., 100, 109
Ellerman, F., 168, 182
Elliot, H., 108
Elsasser, W. M., 128, 141
Elwert, G., 160, 178, 182
Elyasberg, P. E., 117
Eneev, T. M., 36, 38, 47, 55, 109, 117

Fa-Hsien, 15
Farley, M., 107
Fejer, J. A., 108
Ferraro, V. C. A., 75, 108
Firor, J., 157
Fisher, R. A., 132, 141
Fleet, J. F., 15
Frank, L. A., 110
Freden, S. C., 106, 108
Freedman, C., 141
Friedman, H., 69, 70, 108, 160, 178, 182, 183

Galaxy, period-distribution of Cepheids in the, 188
Gaubil, A., 14, 16, 25, 27
Gellman, H., 141
Geomagnetic field in Europe, 127–142
Geomagnetic secular variations, 127–142
Giovanelli, R. G., 167, 173, 183
Gnomons, 3
Gottlieb, M. B., 72, 109
Green, R., 135, 141
Griffith, D. H., 134, 141
Gringauz, K. I., 84, 108
Groves, G. V., 44, 49, 50, 97, 99, 107, 108
Gurevich, L., 152, 183

Hagen, J. P., 108
Hanel, R., 110
Harland, W. B., 132, 141
Heezen, B. C., 135, 141
Hellmann, G., 141
Hendrix, W. C., 142
Henize, K. G., 118, 126
Heppner, J. P., 76, 108
Hertz, H. G., 100, 109
Hipparchus, 19
Hok, G., 84, 108
Ho Ping-Yü, 3, 23
Horowitz, R., 66, 97, 99, 109
Howard, R., 168, 183
Hsüan-Chuang, 15
Hydromagnetic waves, 149
Hynek, J. A., 60, 110

I-Hsing, 3, 4, 5, 6, 7, 12, 13, 14, 15, 17, 20, 22, 23, 24, 25, 26, 27
I-Hsing's and Nankung Yüeh's geodetic expeditions, 4
Imyanitov, I. M., 86, 108
Ionosphere, studied with artificial satellites, 78, 101
Ising, G., 134, 141

Isogonic lines, 138, 139
Isometric presentation, relating variations in magnetic declination, inclination, and intensity, 140
Istomin, V. S., 103, 108

Jacchia, L. G., 97, 108
Jaeger, J. C., 135, 141
Jastrow, R., 39, 40, 108, 117
Jeffreys, H., 53, 108
Johnson, E. A., 134, 135, 141
Jones, L. M., 67, 108
Joy, W. R., 101, 108
Juan Yuan, 5

Katz, L., 108
Keen, M. J., 135, 141
Kiepenheuer, K. O., 143, 144, 145, 153, 163, 170, 172, 183
Khai-Yuan, 10, 12
Khang-Hsi, 27
King, R. F., 134, 141
King-Hele, D. G., 47, 55, 56, 98, 99, 100, 107, 108
Kippenhahn, R., 162, 183
Kiseley, A. A., 121, 126
Kitchen, F. A., 101, 109
Klüber, H. von, 148, 182
Ko Hung, 25
Kozai, Y., 100, 109
Krassovskii, V. I., 105, 109
Kraus, J. D., 61, 62, 109
Krause, H. C. L., 117
Kreplin, R. W., 178, 182
Kundu, M. R., 160, 162, 182, 183
Kuo Shou-Ching, 9, 22
Kupperian, J. E., 29, 69, 70, 108, 178, 182
Kurnosova, L. A., 109
Kuz'min, A. D., 183

Lagow, H. E., 66, 97, 99, 109
Lange, I., 142
Laplace, P. S., 22
Laporte, L., 142
Laughton, A. S., 135, 141
Lawden, D. F., 109
Lebedinskii, A., 152, 183
Leighton, R. B., 151, 183
Leslie, D. C. M., 47, 108
Li Chih-Chhang, 14
Lidov, M. L., 117
Lindsay, J. C., 108
Lines, A. W., 109
Liu Chhuo, 4
Liu Fang, 5, 6, 7
Liu Hsü, 5
Li Shun-Fêng, 6, 20
Loughhead, R. E., 170, 183
Lovell, A. C. B., 95, 107
Lozinskii, A. M., 118, 119, 126
Ludwig, G. H., 107, 110
Lu Gwei-Djen, 3

Magellanic Clouds, 184

Magnetic declinations, 136
Magnetic field of a solar centre of activity, 146
Magnetic measurements from artificial satellites, 74
Magneto-ionic theory, 79
Makita, M., 147, 183
Marchant, M., 100, 109
Martyn, D. F., 76, 109
Masevich, A. G., 118, 119, 126
Massey, H. S. W., 29, 108
Mathewson, D. S., 156, 182
Matjagin, V. C., 122, 126
Ma Tuan-Lin, 23
McCabe, M. K., 170, 173, 183
McIlwain, C. F., 107, 110
McNish, A. G., 134, 141
Menelaus of Alexandria, 19
Mercure, S., 150, 151, 183
Meredith, L. H., 72, 109
Meridian-line, 3
Merson, R. H., 99, 108
Meteorological studies with an artificial satellite, 66
Metric system, 3, 27
Micrometeorites, 86
Mieli, A., 19
Mikhnevich, V. V., 108
Miller, S. C., 150, 183
Milligan, J. C., 109
Mitra, S. K., 47, 109
Mori, S., 15
Moulton, F. R., 41, 109
Murphy, T., 135, 141
Mustel, E. R., 181, 183

Nankung Yüeh, 4, 6, 9, 10, 12, 14, 20, 24, 25
Nazarova, I. N., 103, 109
Needham, J., 3, 4, 27
Neher, H. V., 73, 109
Newkirk, G., 160, 173, 183
Nixon, J., 141
Nonweiler, T., 97, 109
Nordberg, W., 110
Norman, R., 141

Obliquity of the ecliptic, 20
O'Keefe, J. A., 100, 109
Okhotsimskii, D. E., 36, 38, 47, 55, 109, 117

Packer, D. M., 151, 183
Panayotov, L. A., 123, 126
Parker, E. N., 181, 183
Parkyn, D. G., 47, 109
Payne-Gaposchkin, C., 184
Pearse, C. A., 39, 40, 108, 117
Period-luminosity relation, 184–189
Photography of faint artificial satellites, 121
Piddington, J. H., 149, 183
Piggot, W. R., 179, 183
Pikelner, S. B., 181, 183
Plimmer, R. N. A., 100, 109
Priester, W., 97, 109
Ptolemy, 19
Pulleyblank, E. G., 3, 6,
Purcell, J. D., 151, 183
Pythagoras, 23

Radiation belts of the Earth, 105
Radio reflections from ionization clouds of artificial satellites, 61
Ratcliffe, J. A., 79, 109
Ray, E. C., 110
Rense, W. A., 150, 183
Repnev, A. I., 108
Roberson, R. E., 55, 56
Roberts, J. A., 170, 183
Rozhkovskii, D. A., 122, 126

Sadler, D. H., 109
Salomonovich, A. E., 183
Sarton, G., 19
Sawyer, C., 189
Schach, M., 63, 108
Schatzman, E., 153, 183
Schilling, G. F., 98, 109
Scott, J. M. C., 109
Scott, W. E., 142
Setty, S. C., 109
Severny, A. B., 166, 167, 168, 169, 170, 183
Shapiro, B. S., 102, 108
Shapley, H., 184, 187, 188
Sheridan, K. V., 173, 183
Schlüter, A., 162, 183
Schmerling, E. R., 109
Shearman, E. D. R., 101, 108
Short-wave radiations from Sun and other sources, studied by artificial satellites, 68
Shvidkovskii, E. G., 108
Sicinski, H. S., 84, 108
Singer, S. F., 105, 109
Siry, J. W., 36, 109
Smart, W. M., 41, 109
Smith, D. E., 19
Solar centres of activity, 143–183, 179
Solar corona: activity regions and condensations, 153, 162, 175
Solar magnetic fields and solar variability, 152
Solar flares and associated dynamic phenomena, 164
Spencer, A. W., 84, 108
Spitzer, L., 55, 66, 109
Squires, R. K., 100, 109
Stampel, R., 110
Stepanov, V. E., 165, 183
Sterck, J. Damen, 182
Sterne, T. E., 98, 109
Störmer, C., 106, 109
Stroud, W. G., 110
Stuart, F., 150, 183
Su Mien, 6, 7
Suomi, V. E., 66, 110
Sweet, P. A., 147, 183

Tabor, H., 66, 110
Ta-Hsiang, 10, 12, 14, 25
Takakura, T., 178, 183
Tan, Duke of Chou, 9

Taratynova, G. P., 47, 48, 55, 56, 109, 110, 117
Thellier, E., 130, 140, 141
Thellier, O., 140, 142
Thomas, A., (An To), 3, 27
Thomas, J. O., 109
Thomas, R. N., 151, 182
Thompson, G. I., 3
Ting Wên-Chiang, 21
Torrenson, O. W., 135, 141
Tousey, R., 151, 183
Tracking of artificial satellites, 57
Tracking of artificial satellites, radio tracking, 58, 101
Tracking of artificial satellites, optical tracking, 60
Tracking of artificial satellites, radar tracking, 61
Trent, G. H., 173, 183
Tsap, T. T., 181, 183
Tung Tso-Pin, 8, 20, 24

Udalt'sov, V. A., 183
Unsöld, A., 153, 183

van Allen, J. A., 72, 73, 74, 104, 105, 107, 109, 110
van Bemmelen, W., 137, 138, 139, 142
van de Hulst, H. C., 87, 110, 149, 180, 183
van der Waerden, B. L., 19
Vanguard test-spheres, 94
Varāha-Mihira, 19, 20
Vernov, S. N., 105, 106, 110
Vestine, E., 128, 142
Vitense, E., 148, 153, 183
Vitkevich, V. V., 154, 183
Vost, W., 15

Waldmeier, M., 161, 177, 180, 183
Waley, A., 14
Wang Chhung, 25
Wang Phu, 6
Warwick, J. W., 101, 110
Watanabe, T., 135, 142
Weekes, K. G., 81, 101, 107, 110
Weller, F., 15
Weisfeld, M., 108
Wheelon, A. D., 108
Whipple, F. L., 49, 60, 108, 110
White, R. S., 106, 108
Whitney, C. A., 110
Wild, J. P., 173, 178, 183
Willmore, A. P., 61, 110
Wilson, J. G., 108
Wright, A. E., 134, 141
Wu Chhêng-Lo, 24
Wu Tsung, 6
Wu Wang, 9

Yabuuchi Kiyoshi, 20, 22
Yatsunskii, I. M., 110, 117
Yuan-Thai, 10, 12, 14, 25

Zanstra, H., 162, 183
Zelikman, M. Ch., 84, 108
Zero-point of mass-luminosity relation, 184
Zirker, J. B., 49, 108

VOLUME 4

ASTRONAUTICAL HISTORY
CELESTIAL MECHANICS
ARTIFICIAL SATELLITES
GEOPHYSICS
SOLAR RESEARCH
ASTROPHYSICS

L.
61

VOL
4
196

An 8th-Century Meridian Line: I-HSING's Chain of Gnomons and the Pre-history of the Metric System

A. Beer
The Observatories, Cambridge

Ho Ping-Yü
University of Malaya, Singapore

Lu Gwei-Djen
Newnham College, Cambridge

J. Needham, F.R.S.
Gonville and Caius College, Cambridge

E. G. Pulleyblank
Downing College, Cambridge

G. I. Thompson
The Observatories, Cambridge*

Summary

A study is here made of one of the most remarkable pieces of organized field research in the early middle ages — the meridian survey directed by a Buddhist monk I-Hsing and an official astronomer Nankung Yüeh in +725. At a central chain of four stations in Eastern China, at measured distances apart, covering some 200 km., observations were systematically made. It is believed that these consisted of the measurement of solstitial and equinoctial Sun shadows and of polar altitudes. The observations were also carried out at a chain of five further stations, about 2500 km in length in all, from Indo-China to the southern border of Mongolia. A single northernmost station in the vicinity of Lake Baikal had also been the scene of similar observations, thus making it possible to consider an arc of no less than 3800 km. length. The ratio of terrestrial distance units (*li*) to the degree, which it was one of the objects of the survey to ascertain, fixed a civil unit in a manner prefiguring the metric system of a thousand years later. The fact that I-Hsing, Nankung Yüeh and their colleagues accepted this ratio as a constant may imply that some of them envisaged the Earth's sphericity — a view in harmony with some of the ancient schools of Chinese cosmology, but not generally accepted by scholars in their time. The *li* which these astronomers desired to express in terms of the degree appears to have been one of the two usual distance units of the period, namely the "normal short Thang li". A surprising result is that I-Hsing appears to have had remarkably accurate trigonometric tables at his disposal. The paper concludes with a study of the astronomical standardization of the *li* by Antoine Thomas S.J. in 1702 at the request of the Khang-Hsi emperor, nearly a century before the similar standardization of the metre in Europe.

1. Introduction

The fundamental significance of the introduction of the metric system lies in the fact that it was the first great attempt to define terrestrial units in terms of an unvarying cosmical quantity. Scientific lexicographers say that this system took its origin in the need imposed by the development of scientific thought for immutable and at the same time conveniently related units of physical measure. And they imply that this need was not satisfied until the last decade of the +18th century. This may be true enough for Europe but, as will be mentioned in the concluding section of this article, an approach was made to such an immutable unit in China in the first decade of

* Now at the Royal Observatory, Edinburgh.

that century. Moreover, as in the case of many post-Renaissance developments, there were earlier historical aspects of this celestial-terrestrial bond; indeed, we can find already in the +8th century in China a large-scale attempt to establish it.

A great step forward was taken when the idea arose of fixing terrestrial length-measures in terms of astronomical units. That this could have occurred to the scholars at all was due to the fact that the Sun's shadow thrown by an 8-foot gnomon at Summer Solstice had a very convenient length (about 1·5 ft) at the latitude of Yang-chhêng, the "centre of the Central Land". From ancient times use had been made of the "gnomon shadow template" (*thu kuei*), a standard rule made of pottery, terra-cotta, or jade, equivalent in length to the solstitial shadow. This was used for the determination of the exact date of the solstice each year.[1]

It was a long-standing idea that the shadow-length increased one inch for every thousand *li* to the North of the "Earth's centre" at Yang-chhêng, and decreased in the same proportion as one went to the South. After the end of the Han period (+3rd century), measurements made as far south as Indo-China soon disproved this numerically, but it was not until the Thang Dynasty (+8th century) that a systematic effort was made to cover a great range of latitudes.

This effort aimed at correlating the lengths of terrestrial and celestial measures by finding the number of *li* which corresponded to 1° change in the altitude of the Pole Star (giving the geographic latitude of the observer's position) and thus, in effect, fixing the length of the *li* precisely in terms of the Earth's circumference. The meridian line set up for this purpose takes its place in history between the meridian line of Eratosthenes (ca. −200) and those of the astronomers of the Caliph al-Ma'mūn (ca. +827). Its detailed examination is the object of this paper.

2. First Suggestion of a Survey

When Liu Chhuo, in the early part of the +7th century, realized the fallacy of the statement that a change by 1 inch in the length of the Sun's shadow corresponded to a change of 1000 *li* in distance, he wrote to the Emperor as follows:[2]

"We beg Your Majesty to appoint water-mechanics and mathematicians to select a piece of flat country in Honan and Hopeh, which can be measured over a few hundred *li*, to choose a true North-South line, to determine the time with water-clocks, to [set up gnomons] on flat places [adjusting them with] plumb-lines, to follow seasons, solstices and equinoxes, and to measure the shadow of the Sun [at different places] on the same day. From the differences in these shadow-lengths the distance in *li* can be known. Thus the Heavens and the Earth will not be able to conceal their form and the celestial bodies will not prevent us from knowing their measurements."

His advice was not adopted by the Sui emperor.

3. The Expeditions of I-Hsing and Nankung Yüeh

However, under the following dynasty, the Thang, in the years +721 to 725, the required expeditions were organized under the direction of the Astronomer-Royal, Nankung Yüeh, and a Buddhist monk, I-Hsing, one of the most outstanding

[1] See J. Needham, *Science and Civilization in China* (hereinafter abbreviated as *SCC*), Cambridge Univ. Press, vol. 3, pp. 286 ff, 1954.

[2] *Sui Shu*, ch. 19, p. 20 *a*, *b*; cf. *Chhou Jen Chuan*, ch. 12, pp. 150 ff.

mathematicians and astronomers of his age.[3] Our sources of information about them are quite extensive.[4] First, there is the *Chiu Thang Shu* (Old History of the Thang Dynasty),[5] ch. 35, pp. 6a ff.[6] This was finished in +945. A parallel text is found in the *Hsin Thang Shu* (New History of the Thang Dynasty),[7] completed in +1061 (ch. 31, pp. 5b ff.). Further material is in the *Thang Hui Yao* (Administrative Statutes of the Thang Dynasty),[8] ch. 42, (p. 755), a work dating from +961. This last text often gives values differing from the other two. A very much later text which clarifies some of the data is the *Chhou Jen Chuan* (Lives of the Astronomers and Mathematicians),[9] written by Juan Yuan in +1799, (ch. 16, pp. 202 ff.).

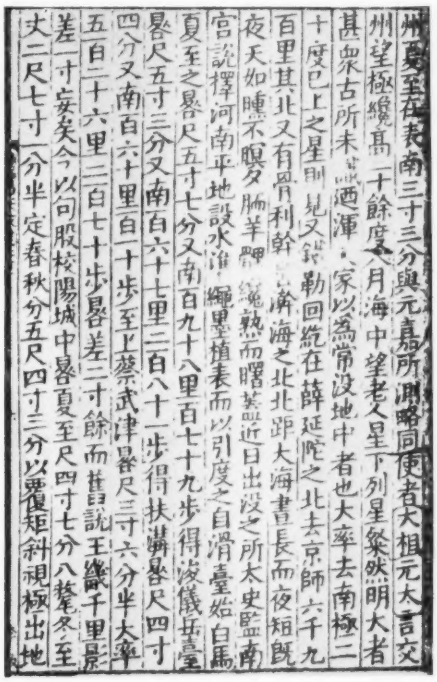


Fig. 1. I-Hsing's expedition as reported in the *Hsin Thang Shu*.

The *Chiu Thang Shu* was compiled in its final form under the editorship of Liu Hsü in the Later Chin dynasty and was, as we have said, completed in +945. It was however based on an existing *Thang Shu* of which the latest version was that of Liu Fang, which came down to about the year +756 with some minor additions a

[3] *SCC*, vol. 3, p. 202.
[4] A preliminary account was given in *SCC*, vol. 3, pp. 292 ff.
[5] Hereinafter abbreviated as *CTS*.
[6] The text is on pp. 6 and 7, but besides this there is an appendix to the chapter on pp. 8b, 9a, which gives values, some variant, in tabulated form. This seems to have been added at some time later than that of the writing of the text. Some of the discrepancies between text and appendix are quite conspicuous, e.g. for station 9 of Table I, where the text gives 20°4 for the polar altitude, while the appendix gives 26°6.
[7] Hereinafter abbreviated as *HTS*.
[8] Hereinafter abbreviated as *THY*.
[9] Hereinafter abbreviated as *CJC*.

little later. There is good evidence that the basic annals and the biographies in the existing *Thang Shu* were simply copied into the *Chiu Thang Shu* with a minimum of editing.[10]

In the case of the monographs it was necessary to complete them individually to the end of the dynasty. Moreover it is not clear from the sources just what monographs already existed in Liu Fang's *Thang Shu*. The evidence has to be considered separately in the case of each one. It is *a priori* probable however that there would have been a monograph on astronomy, since this was one of the monographs regularly included in all standard histories. Moreover when we examine the structure of the existing monograph on astronomy in *CTS* it becomes apparent that it is made up of two parts, (*a*) material arranged chronologically under specific topics down to the year +756, (*b*) additional material down to the reign of Wu Tsung (+841 to +846), the last for which a "veritable record" existed.

The monograph consists of two chapters. The first, which contains the information about the observations of +724 and +725, is an account of the major undertakings of the Bureau of Astronomy. It contains no material later than +756. The second chapter begins with a short preface introducing I-Hsing's revision of a brief work by Li Shun-Fêng on the twelve Jupiter stations[11] which is incorporated textually into the monograph. Then follows a list of eclipses through the dynasty. Down to the year +756 there is only an unbroken list of dates. After that, each entry is accompanied by a record of the eclipse, sometimes only of a few words but sometimes fairly long. Moreover the manner of recording the dates is different. Down to +756, the days of the month are given numerically, thereafter by cyclical signs. Then follow separate lists of heavenly phenomena, comets, planets, meteors, aurorae. In each case the list is chronological from the beginning of the dynasty, there is no record later than +756, and days of the month are recorded numerically. After this comes an unbroken section giving a straight chronological unclassified record of all such phenomena from +756 to +846. Days are given cyclically. The chapter concludes with a short paragraph on the organization of the Bureau of Astronomy, all of which, except for the last entry, is concerned with the period before +756.

It seems clear that the *CTS* editors simply incorporated an existing text, no doubt that in Liu Fang's work, into their book, adding such entries as they found in the various "veritable records" of reigns from +756 to +846. This supposition is quite in accord with what we know about the hasty way in which the *CTS* was compiled.

The *Thang Hui Yao*, like the *CTS*, is a work completed after the end of the dynasty (in +961 by Wang Phu) but incorporating previously existing works. The first of these was the *Hui Yao* of Su Mien, completed in +804; and it is entirely probable that the record of the observations of I-Hsing and Nankung Yüeh in the *Thang Hui Yao* comes from it. This in turn is clearly related to the account in the *CTS*, though very much condensed. The list of data it gives for the various stations is, apart from textual variants, identical with that found in the appendix to the first chapter of the monograph in the *CTS*. Moreover the occurrence of an identical sentence in the two works shows that they were at least based upon the same source. This sentence occurs in the discussion of the results in the *CTS*, which shows that this part, as well as the actual description of the observations and record of data, had been written at least as early as +804.

---

[10] See E. G. Pulleyblank in *Bulletin of the School of Oriental and African Studies*, **13**, pp. 448–73, 1950

[11] See *SCC*, vol. 3, p. 402.

Thus the evidence points overwhelmingly to the conclusion that the whole account as found in the *CTS* already existed in Liu Fang's *Thang Shu*, where it was available to be used by Su Mien, and from which it was copied by the *CTS* editors. It was therefore written up within thirty years or so of the events with which we are concerned.

The sources tell us that at least eleven stations were established where simultaneous measurements of shadow lengths were made, using identical 8-foot gnomons. The latitude of these stations ranged from 17°.4 (at Lin-i, Indrapura in Champa, capital of the Lin-i State, not far from modern Hué) up to 40° (at Wei-chou, an old city near modern Ling-chhiu, near the Great Wall in northern Shansi, and almost at the same latitude as Peking). And there was even another place still farther North,

Fig. 2. Nankung Yüeh's gnomon at the Yang-chhêng Observatory (from Tung Tso-Pin *et al.*)

the country of the Thieh-lo (Tölös) horde of Turkic nomads near Lake Baikal; it is not quite certain, however, whether any actual observations were made there in I-Hsing's time; the figures seem to have been extrapolations. Furthermore, the statement that all of I-Hsing's stations were approximately on a North-South line requires some qualification; see Section 8 below. Most of them were on the great plains North and South of the Yellow River, the last-mentioned one at the northern border of China proper, and the two in the far South in Indo-China.

Station 5 (see Table 1) is of particular interest. Yang-chhêng was a place which had been for centuries the seat of the central imperial observatory of China. Although it did not belong to the measured central chain of stations, it is the only place where

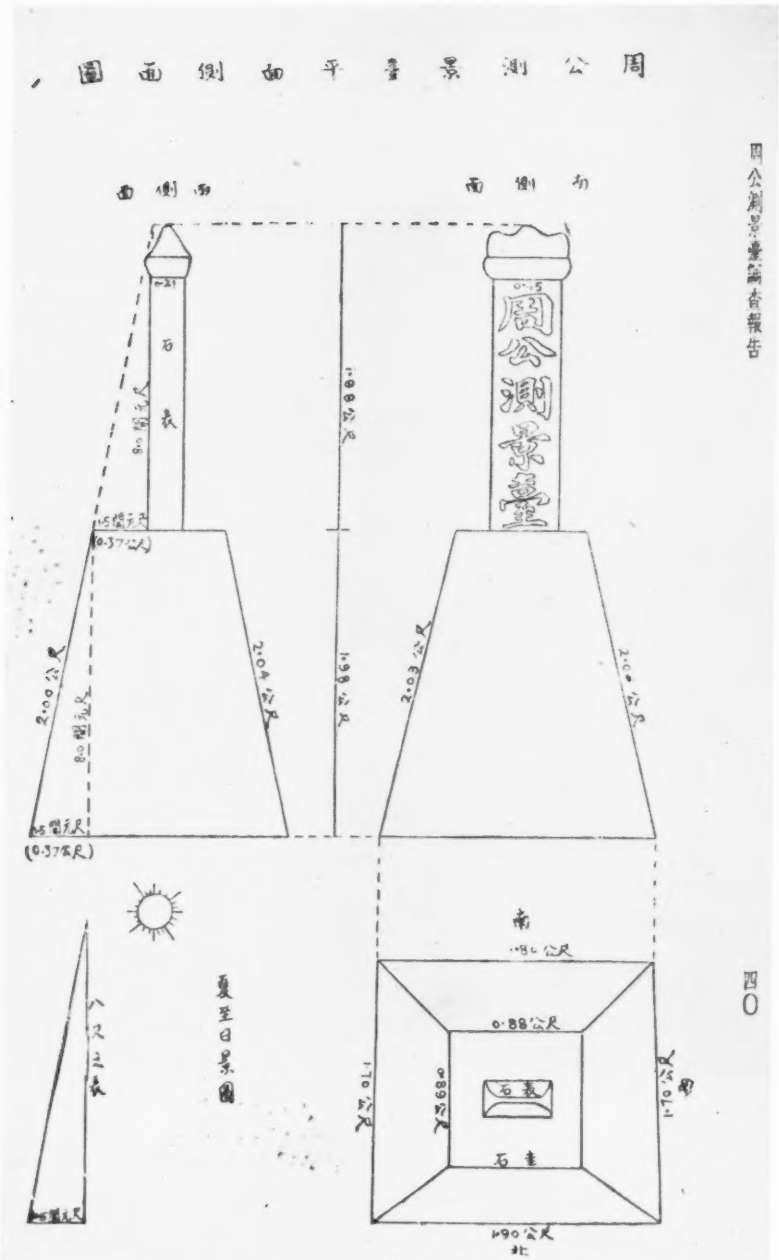


Fig. 3. Scale drawings of Nankung Yüeh's gnomon
(from Tung Tso-Pin *et al.*)

one of the original gnomons of I-Hsing and Nankung Yüeh is still preserved (Figs. 2 & 3).[12] On the South side this bears an inscription terming it "Chou Kung's Tower

[12] Tung Tso-Pin *et al.*, *Chou Kung Tshê-Ying-Thai Tiao-Chha Pao-Kao* (Report on Chou-Kungs' Tower for the Measurement of the Shadow of the Sun). Commercial Press, Chhangsha, 1939, pp. 38, 39, 40, 94.

for the Measurement of Sun Shadows"[13] and it is known to have been erected by the Astronomer-Royal, Nankung Yüeh, in pursuance of an imperial edict in +723. The construction is such that at the Summer Solstice of that time the shadow just extended to the top of the pyramidal base, the slope of the North side of which corresponded exactly with the edge of the shadow. In later times, Kuo Shou-Ching, Astronomer-Royal of the Yuan dynasty, made shadow measurements at Yang-chhêng with a gnomon of 40 feet in height and a measuring-scale of some 120 feet. This was about +1270, and there still remains intact at the place the massive tower and scale constructed in Ming times following his methods.[14]

4. Translations of the Sources[14a]

CHIU THANG SHU, ch. 35, *pp.* 6*a ff.*

"*Sun Shadows.* In the *Chou Li* (Record of Institutions of the Chou Dynasty)[15] unde the heading of the Ta Ssu Thu it says that [this official] uses the method of the gnomon shadow template to measure the 'depth of the Earth' and to establish correctly the sun-shadows in order to find the centre of the Earth. To the East of the Sun the shadow is that of the evening and there is much wind; to the West of the Sun the shadow is that of the morning and there is much overcast sky. The place where the shadow at [Summer] Solstice is 1 ft 5 in. is called the centre of the Earth. This is where Heaven and Earth unite, where the four seasons intermingle, wind and rain join together, where the Yin and the Yang combine. Then all things prosper and the royal territory may be established there. The commentator Chêng [Hsüan][16] considered that the Sun shadow differed by 1 in. for every 1000 *li* on the Earth's surface. From the place where the shadow is 1 ft 5 in. it would be [he thought] 15,000 *li* to the South to the place directly below the Sun. The Earth makes its four excursions and the stars rise and fall within a range of 30,000 *li*, so by taking the half of this one gets the centre of the Earth.

5 *Comm.* Chêng Chung[17] said that the gnomon shadow template should be 1 ft 5 in. On the day of the Summer Solstice an 8 ft gnomon was set up; where the shadow exactly fits this, that is the centre of the Earth. At Yang-chhêng in Yin-chhuan this is so.

"We respectfully submit that in the *Nan Yüeh Chih* (Records of Southern Yüeh) it is recorded that in the Yuan-Chia reign-period of the [Liu] Sung dynasty [+424 to +453] there was an expedition to Lin-i. In the 5th month they set up a gnomon, and when observation was made the Sun was to the North of the gnomon and the shadow was to the South. At Chiao-chou the Sun shadow was seen to be 3 in. [to the South]. At Lin-i it was seen to be 9·1 in. [to the South].[18] This is what is meant by the saying 'opening the Earth's door to face the Sun'. Now Chiao-chou is approximately over 9000 *li* from Lo[-yang], that is, including the twists and turns of the land and sea route. If one does not bring into account the calculation from the gnomon, and only estimates directly the true distance, would it not be about 5000 *li*?

"In the 12th year of the Khai-Yuan reign-period [+724] the Astronomer-Royal

[13] Tan, Duke of Chou, was an outstanding Chinese culture-hero. There is no reason to doubt his historicity as the younger brother of the first High King of the Chou empire, Wu Wang, whose decisive defeat of the Shang State occurred in the close neighbourhood of −1030. Duke Tan was a great patron of the arts and sciences, hence his connection with astronomical observations.

[14] Described in *SCC*, vol. 3, pp. 296 ff.

[14a] The figures enclosed in the square cartouches in the margins below are the identification numbers given by us to the observing stations assembled in Table I, etc.

[15] Perhaps −2nd century. [16] +2nd century. (*Note*: 1 Chinese foot = 10 Chinese inches).

[17] +1st century. [18] The text actually says North, but we correct following *Sui Shu* ch.19, p. 24a.

10 was ordered to make observations at Chiao-chou of the Sun shadow at Summer Solstice, and it was found to be 3·3 in. to the South of the gnomon. This was in good agreement with the observation made in the Yuan-Chia reign-period [+424 to +453]. This being so, if one went South from Yang-chhêng along a road as straight as a bowstring to the point directly below the Sun it would not be as much as 5000 *li*. The Commissioners for Shadow Measurements Ta-Hsiang and Yuan-Thai say that at Chiao-chou if one observes the pole it is elevated above the Earth's surface only a little more than 20°. Looking South in the 8th month from out at sea Lao-jen [Canopus] is remarkably high in the sky. The stars in the heavens below it are very brilliant and there are many large and bright ones which are not recorded on the charts and the names of which are not known. In general all the stars which are more than 20° from the southern pole are all visible. Indeed it is the part of the sky which the ancient Hun Thien [Celestial Sphere] school of astronomers regarded as permanently below the horizon and therefore not to be seen.

1 "It may also be remarked that according to what the historians have recorded of the Chên-Kuan reign-period [+627 to +649] the Uighurs of the Thieh-lo tribes, living North of the Hsüeh-yen-tho people, were 6900 *li* North of the capital [Chhang-an]. Moreover there are the Ku-li-kan people who live to the North of the Uighurs, dwelling North of Han-hai [Lake Baikal] where grass is plentiful and there are many herbs, and where fine horses are produced capable of going several hundred *li* [in a day]. To the North of that there is still some distance to the Great Sea [the Arctic]. The days are long and the nights short. After the Sun has set the sky is still half lit and if you cook a sheep the outside is hardly done before the dawn begins to appear in the East. For it is close to the place where the Sun goes down and emerges. Both these matters [concerning the far South and the far North] have not previously been recorded in books.

"In the 13th year of the Khai-Yuan reign-period [+725] Nankung Yüeh the Astronomer-Royal selected a region of level ground in Honan and using water-levels and
3 plumb-lines set up 8 ft gnomons with which he made measurements. Beginning at Pai-ma Hsien in Hua-chou he found the Summer Solstice shadow to be 1 ft 5·7 in. Proceeding South from the observation station at Hua-chou 198 *li* 179 *pu* they
4 reached the old observation station at Chün-i in Pien-chou with its gnomon; there the Summer Solstice shadow was found to be 1 ft 5 in. *wei-chhiang*.[19] Again, going South from Chün-i for 167 *li* 281 *pu* they reached the gnomon at Fu-khou Hsien in Hsü-
6 chou, which gave a length of 1 ft 4·4 in. at Summer Solstice. Then 160 *li* 110 *pu* South of Fu-khou there was another gnomon at Wu-chin near Shang-thai in the district of Yü-chou, which gave a shadow of 1 ft 3·65 in. at Summer Solstice. In all,
7 therefore, in a distance of 526 *li* 270 *pu*, the difference in shadow length was a little more than 2 in. This completely disagreed with the opinion of the ancient scholars that for a distance of 1000 *li* in the royal territory there would be a variation of 1 in. in the shadow's length.

"Now if one makes comparisons with the aid of the 'right-angled triangle diagram',
5 the Summer Solstice shadow at Yang-chhêng [being] 1 ft 4·8 in. *jo*,[20] the shadow length at Winter Solstice [will be] 12 ft 7·15 in., and at the Spring and Autumn

---

[19] *Wei-chhiang* probably means *shao-chhiang* or 5/16; the length would then become 1 ft 5·312 in.
[20] *Jo* is −1/8, hence the value is 1 ft 4·788 in.

Equinoxes 5 ft 4·3 in. Now if you use the 'slope of the inverted square [diagram]' to estimate the polar altitude, it is 34·4 *tu*,[21] (at Yang-chhêng).

Comm. Divisions of *tu* follow the decimal system.

3 If you estimate it for the gnomon at Hua-chou, the elevation is 35.3 *tu*.

Comm. The difference from Yang-chhêng is 0·9 *tu*.

4 If you estimate it for the gnomon at Chün-i, the elevation is 34·8 *tu*.

Comm. The difference from Yang-chhêng is 0·9 *tu*.

7 If you estimate it for the gnomon at Wu-chin, the elevation is 33·8 *tu*.

Comm. The difference from Yang-chhêng is 0·9 *tu*.

"Although there may be small errors of excess and defect in the fractions which are difficult to discern with the eye, yet over the whole distance of 526 *li* 270 *pu* there is a difference of 1·5 *tu* in polar elevation, and therefore 351 *li* 80 *pu* correspond to a change of 1 *tu*. According to the variations in distance from the celestial pole there is change in the path of the ecliptic and the [lengths of the] shadow cast [by the Sun] on it.

"Using this value just derived as the ratio, we can extend it. For successive years

8 shadows at Lang-chou have been measured. At the Summer Solstice it is 0 ft 7·7 in., at the Winter Solstice 10 ft 5·3 in., and at the equinoxes 4 ft 3·75 in.

Comm. Measured by the diagram, the equinoctial shadow is 4 ft 4·7 in.

If you estimate the polar altitude according to the slope on the 'diagram', the elevation is 29·5 *tu*.

Comm. The difference from Yang-chhêng is 5·2 *tu*.

2 At the Hêng-yeh army post in the district of Wei-chou,[22] the shadow at Summer Solstice has been measured and found to be 2 ft 2·9 in.

At Winter Solstice it is 15 ft 8·9 in. and at the equinoxes it is 6 ft 4·45 in.

Comm. Measured by the diagram, the equinoctial shadow is 6 ft. 6·35 in.

If you estimate the polar altitude according to the slope on the 'diagram', the elevation is 40 *tu*.

Comm. The difference from Yang-chhêng is 5·2 *tu*.

The difference in these two polar altitudes is thus $10\frac{1}{2}$ *tu*, and the distance between the stations [will be] 3680 *li* 90 *pu*.

Comm. From Yang-chhêng to Lang-chou the distance [will be] 1826 *li* 196 *pu*. That from Yang-chhêng to the Hêng-yeh army post in Wei-chou [will be] 1861 *li* 214 *pu*.

The difference in the length of Summer Solstice shadow is thus 1 ft 5·3 in.

Comm. From Yang-chhêng to Lang-chou the difference is 0 ft. 7·2 in. and from Yang-chhêng to the Hêng-yeh army post 0 ft 8 in.

The difference in the length of Winter Solstice shadows is thus 5 ft 3·6 in.

Comm. From Yang-chhêng to Lang-chou the difference is 2 ft. 1·8 in. and from Yang-chhêng to the Hêng-yeh army post 3 ft 1·8 in.

The differences between Summer Solstice shadows and all shadows in the South are small, those between Winter Solstice shadows and all shadows in the North are large.

"Furthermore, if we make comparisons with the 'diagram', at An-nan (Chiao-

9 chou) the Sun at Summer Solstice is 2·4 *tu* north of the zenith, and the polar elevation 20·4 *tu*. The Winter Solstice shadow is 7 ft 9·4 in.

Comm. The equinoctial shadow is 2 ft 9·3 in.

[21] *Tu* is the Chinese degree. 1 *tu* = (360/365·25)°, i.e. 0°98563.

[22] The correct reading of this place-name is Yü-chou, but to avoid confusion with one of the other stations, we retain the normal pronunciation of the character.

The difference from Yang-chhêng is 14·3 *tu*, and the distance [will be] 5023 *li*.

"Coming to Lin-i, [according to the] diagram, the Sun is [at Summer Solstice] 6·6 *tu chhiang*[23] north of the zenith. The polar elevation is 17·4 *tu*, and the circle of perpetual visibility amounts to 35 *tu*. The length of the shadow [at Winter Solstice] is 6 ft 9 in., and the distance from Yang-chhêng [will be] 6112 *li*.

"Let us suppose that going north from Yang-chhêng to the country of the Thieh-lo [people] the difference is also 17·4 *tu* [in polar elevation], being exactly the same as that between Yang-chhêng and Lin-i. Then in the 5th month (at the Summer Solstice) the Sun will be 27·4 *tu* [south of] the zenith, the polar elevation will be 52 *tu*, and the circle of perpetual visibility 104 *tu*. The shadow at Summer Solstice will be 4 ft 1·3 in., at the Winter Solstice 29 ft 2·6 in.

Comm. And at the equinoxes 9 ft 8·7 in.

P. 8*b* (*Conclusion*).

"Thus the differences in the Sun shadows vary as between the Winter and Summer Solstices and also as between northern and southern latitudes. But the ancient scholars equalized the differences everywhere with a fixed value in terms of *li*, and so failed to give a true account. Accordingly the monk I-Hsing prepared the 'Ta Yen' diagram and also the 'inverted square diagram' covering the range from the furthest South to the furthest North. He also made twenty-four diagrams in order to investigate the computations of the eclipses of the Sun, and to establish the lengths of the indicator-rods of the night clepsydra. Here we record the shadow lengths in feet and inches of all the observation centres."

Appendix, tabulated

No.	Station	Pol. Alt. *tu*	Sum. Solst.	Wint. Solst.	Equinoxes	
			ft in.	ft in.	ft. in.	
10	Lin-i	17·4	−5 7	+ 6 9	+2 8·5	
9	An-nan	26·6	−0 3·3	+ 7 9·4	+2 9·3	= Chiao-chou
8	Lang-chou	29·5	+0 7·7	+10 5·3	+0 4·7	= Wu-ling
7a	Hsiang-chou				+4 8	
7	Yü-chou	33·8	+1 3·65	+12 3·8	+5 2·8	= Wu-chin nr. Shang-tshai
6	Hsü-chou	34·3	+5 4·4	+12 5·3	+5 3·7	= Fu-khou
4	Pien-chou	34·8	+1 5·3	+12 8·5	+5 5	= Chün-i (Khai-fêng)
3	Hua-chou	35·3	+1 5·7	+13	5 5·6	= Pai-ma Hsiên
2a	Thai-yuan				+6	
2	Wei-chou	30·0	+2 2·9	+15 8·9	+6 6·3	= Hêng-yeh army post

THANG HUI YAO, *ch.* 42 (*p.* 755)

"In the 12th year of the Khai-Yuan reign-period [+724] on the 23rd day of the 4th month, a command was issued to the Astronomer-Royal, Nankung Yüeh, and the officials of the Bureau of Astronomy Ta-Hsiang and Yuan-Thai, to proceed by the post-station routes to An-nan, Lang-chou, Tshai-chou, Wei-chou etc. and to measure the lengths of the Sun shadows, a report to be made on the day of their return. For several years they took observations and when they came back to the capital they compared them in conference with the monk I-Hsing. . . At Lang-chou,

[32] *Chhiang* is +1/8; therefore the value is 6·6125 *tu*.

Hsiang-chou, Tshai-chou, Hsü-chou, Honan-fu, Hua-chou, Thai-yuan, etc. there were also in each case Commissioners, and all brought back various results.

"On the basis of the northern and southern Sun shadows, I-Hsing made comparisons and estimates. He used the 'right-angle triangle' method to calculate them. Roughly speaking, the distance between the North and South poles was found to be scarcely more than 800,000 *li*."

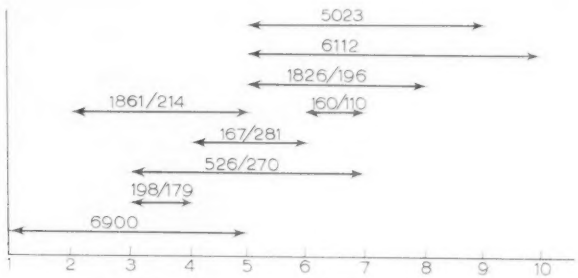


Fig. 4. I-Hsing's stations. Distances in *li* and *pu*.

Figure 4 illustrates the relative distances, while Table 1 summarizes the measurements as follows:

Columns (1) Serial number of station
(2) geographical location
(3) polar altitude in Chinese degrees (*tu*)
(4) shadow length at Winter Solstice, WS,
(5) shadow length at Equinoxes, ES
(6) shadow length at Summer Solstice, SS.

Table 1. Assembly of Data

Station	Place	Pol. Alt.	WS	ES	SS
1	Thieh-lo	**52**[24]	**29·26**	*9·87*[25]	**4·13**
2	Wei-chou	**40**[26]	**15·89**	**6·445**‡[27] *6·635**	**2·29**†
2a	Thai-yuan			6·0‡[28]	
3	Hua-chou	**35·3**	13	5·56	**1·57**†
4	Pien-chou	**34·8**	12·85	5·5	**1·53**‡[29]
5	Yang-chhêng	34·7[30]	**12·715**[31]	**5·43**[32]	**1·4788**†[33]
6	Hsü-chou	34·3	12·55[34]	5·37	**1·44**‡[35]
7	Yü-chou	**33·8**	12·38	5·28	**1·365**†[36]
7a	Hsiang-chou			4·8†[28]	
8	Lang-chou	**29·5**	**10·53**	**4·375**‡[37] *4·47**	**0·77**†
9	Chiao-chou	**20·4**[38] **21·6**	**7·94**	*2·93*[39]	−0·33†
10	Lin-i	**17·4**	**6·9**	2·85	−0·57†[40] −**0·91**†

Remarks
* Calculated value; clearly stated.
† Values from the text clearly observational.
‡ Values probably observational.
All remaining figures are implied computed.

CTS text in bold type
CTS commentary in italic type
CTS appendix in normal type

[24] According to the texts this figure is based on a supposition that the difference in polar altitude between Yang-chhêng and the Thieh-lo country is the same as that between Yang-chhêng and Lin-i, i.e. 17·4 *tu*. This would make the polar altitude 34·7+17·4 = 52·1 *tu*. The figure 52 is evidently a rounding off to the nearest *tu*.
Shadow lengths are consistent with a value of 51·7 *tu*.

5. Results

The results of these expeditions were that the terrestrial distance corresponding to 1° was determined to be 351 *li* 80 *pu*; and the difference in shadow length was found to be very close to 4 inches for each 1000 *li*, or four times the amount accepted by the "scholars of former times". Additional details are given in Gaubil's tabulation, who wrote[41]: "Quand le Bonze Y-Hang n'auroit fait autre chôse que de procurer tant d'observations de la hauteur du Pôle, et de déterminer la grandeur du Ly en le rapportant aux degrés de latitude, on lui auroit toujours une obligation infinie." While the length of the central series of stations was some 150–215 km., that of the whole series was as much as 2500 km., and if the most northerly place is included, a line of no less than 3800 km. was considered. This work must therefore surely be regarded as the most remarkable piece of organized field research carried out anywhere in the early Middle Ages. Even if the ground distances of the furthest stations were not measured, there can be no doubt that observations of Sun shadow lengths were systematically made at them. The chief observers, Ta-Hsiang and Yuan-Thai (probably monks and trained by I-Hsing himself), were also responsible for a special expedition to map the southern constellations down to within 20° of the southern celestial pole.[42]

Five centuries later, in +1221, the work of I-Hsing's survey was completed by the Taoist Chhiu Chhang-Chhun and his party, who made gnomon observations at the Summer Solstice on the bank of the Kerulen river in Northern Mongolia (about 48°N), when travelling on a visit to Chingiz-Khan at Samarqand. This incident is recorded by Li Chih-Chhang in his *Chhang-Chhun Chen Jen Hsi Yu Chi*.[43]

---

[25] *HTS* has 5·87, an obvious error.
[26] *CTS* appendix has 30 *tu*, an obvious error.
[27] These are the figures given in the text of the *CTS*. *HTS* has 6·625 for the calculated value. *CTS* appendix and *THY* have 6·63 and 6·62 respectively, evidently representing the calculated rather than the observed value.
[28] These values occur only in *CTS* appendix and *THY*.
[29] This value is given in *CTS* appendix, *THY* and *HTS*. *CTS* text has 1·5 *wei-chhiang*, i.e. 1·5325.
[30] *THY* only. *CTS* text and *HTS* have 34·4. 34·7 is confirmed by the fact that it is the value that was evidently used in calculating distances. It is also the value most consistent with the differences in polar altitudes given in the texts.
[31] *THY* has 12·71.
[32] *THY* has 5·45.
[33] *HTS* has 1·478; *THY* has 1·49.
[34] *CTS* appendix has 12·53; *THY* has 12·5. Except for SS, figures for Hsü-chou are not given in *CTS* text.
[35] *CTS* appendix has 5·44, an obvious error.
[36] *THY* has 1·36.
[37] *CTS* appendix has 0·47, an obvious error. *THY* has only the calculated value, 4·47.
[38] This is the value of *CTS* text and *HTS*. It is the one used for calculating the distance from Yang-chhéng. On the other hand the value 21·6 given by *THY* agrees with the statement in the *CTS* text and *HTS* that at midsummer the Sun is 2·4 *tu* North of the zenith (21·6 + 2·4 = 24 *tu*, the value used for the obliquity of the ecliptic). *CTS* appendix has 26·6, an obvious corruption of this latter figure. The shadow measurements for SS and WS are also consistent with the value 21·6 *tu* but the ES value requires 20·4. The figure 20·4 may well have arisen through confusion with the contiguous figure of 2·4 *tu* for the distance of the Sun North of the zenith.
[39] *THY* has 2·65.
[40] −0·57 is given in *CTS* appendix, *HTS* and *THY*. This is consistent with ES = 2·85. −0·91 is given at the beginning in the *CTS* text and *HTS*. This is consistent with Pol. Alt. = 17·4, WS = 6·9. The value was an observed one obtained in the Liu Sung period.
[41] A. Gaubil, *Histoire Abrégée de l'Astronomie Chinoise*, vol. 2, p. 78, 1732.
[42] Cf. *SCC*, vol. 3, p. 274.
[43] Ch. 1, p. 10a. See p. 66 in A. Waley's translation: *The Travels of an Alchemist*. Routledge, London, 1931.

It appears established (see Table 4) that I-Hsing used only four stations to determine directly (by measuring on the ground) the distances from each other, expressing these in units which he called *li*.

The two possible values of the *li* which may be considered in this context are as follows:

Table 2. Thang li

li	metres per *li*.	*li* per km.	*li* per degree
Long Thang *li* (foot of 0·2957 m.; 6 ft. = 1 *pu*; 300 *pu* = 1 *li*)	532	1·88	208
Normal short Thang *li* (foot of 0·2456 m.; 6 ft. = 1 *pu*; 300 *pu* = 1 *li*)	442	2·26	250

These values of metres per *li* agree closely with those derived by Mori Shikazō from extant Thang footrules.[44] Values for the *li* as low as 193 m. seem to have been used in the Thang and earlier by Buddhist pilgrims, with some relation to the Indian *yōjana*.[45]

The recorded distances were:

3–4 = 198 *li* 179 *pu*[46]
4–6 = 167 *li* 281 *pu*
6–7 = 160 *li* 110 *pu*

Combining these data with Table 1 we get the following:

Table 3. Li-*degree relation*

(1) Pair of stations	(2) Difference in Polar Altitudes from Table I, col. 3	(3) Distance in *li*	(4) *li* per degree, from (2) & (3)
3–4	0·5	198·6	397·2
4–6	0·5	167·9	335·8
6–7	0·5	160·4	320·8
3–7	1·5	526·9	351·3

If we attempt to obtain from these data a relation between the length of a meridian arc of 1° and the unit-length *li*, we have 1° = 526·9/1·5 = 351·267 *li* = 351 *li* 80 *pu*. This was in fact I-Hsing's main conclusion. As Chu Kho-Chen has said:[47]

"Nankung Yüeh, the Astronomer-Royal, selected a piece of flat ground in Honan and measured the distances using a water-level and ropes. He began with Hua-chou in the north of the Yellow River, passed Pien-chou and Hsü-chou and reached Yü-chou. He measured the latitudes of the four places of Pai-ma, Khai-fêng, Fu-khou and Shang-tshai and came to the conclusion that one degree along the meridian was equivalent to 351 *li* 80 *pu* (during the Thang Dynasty: 300 *pu* = 1 *li*)."

The considerable disagreement between the *li*/degree value for the first pair and

[44] S. Mori "Kan Tō Ichiri no Nagasa" (On the length of the *li* in the Han and Thang periods), *Tōyōshi Kenkyū*, **5**, 6, p. 438, 1940.

[45] See W. Vost "The Lineal Measures of Fa-Hsien and Hsüan-Chuang" *J. R. Asiat. Soc.* 1903, p. 65, J. F. Fleet, "The *Yōjana* and the *Li*", *J. R. Asiat. Soc*, **38**, 1011, 1906; and pp. 229 and 462, 1912; and F. Weller, "*Yōjana* und *Li*, bei Fa-Hsien," *Z. dtsch. morgenländ. Ges.* **74**, 225, 1920.

[46] The *li* = 300 *pu*. See footnote 66.

[47] "Chung-Kuo Ku-Tai tsai Thien-Wên-Hsüeh shang Wei-Ta-ti' Kung-Hsien" in *Wo-mên Wei-Tati Tsu Kuo*, ed. anon. Peking, 1951, p. 62.

those for the other two led d'Anville in 1761 to the conjecture that the distance measurement for the first pair of stations (198·6) was alone aberrant.[48] Lacking access to the Chinese texts, he attributed it to a printing error of Gaubil for 168·6; but in fact not only the individual value, but also the total sum of the whole set of three values (i.e. the distance between the stations 3 and 7) is unmistakably recorded in the texts.

Apart from the use of the above relation, there is no other way of relating the *li* of the survey with other units. We may, however, refer to our knowledge of the various Thang *li* from other sources. As the distance between Hua-chou and Pien-chou is given in Thang gazetteers[49] as 210 *li*, the agreement with the survey-measurement of 198·6 *li* suggests in itself that the normal short *li* was being used; this may be supported by other distances in the gazetteers.

6. Distance Measurements

We now proceed to show that all other ground distances between the more distant observing stations, as shown in Fig. 4, were actually calculated and not observed values, obtained by making use of the relation $1° = 351\ li\ 80\ pu$ (351·267 *li*), and of the recorded polar altitudes p (Table 1).

This is demonstrated in Table 4.

Table 4. Use of the relation in computing distances

Station serial no.	Pol.Alt. *tu*	Differences △ *tu*	Distances = △ × 351·267 in *li*/*pu*	Distances (*CTS*)
2	40°0			
5	34·7	2–5 = 5·3	1861/214	1861/214
8	29·5	5–8 = 5·2	1826/176	1826/196
—	—	2–8 = 10·5	3688/93	3688/110
9	20·4	5–9 = 14·3	5023·1	5023
10	17·4	5–10 = 17·3	6077·0	6112[50]

A cursory examination of the scale of the celebrated *Yü Chi Thu* map[51] carved in stone in +1137 suggests that the value $1° = 351\ li$ may have been used later for marking certain long distances in the cartography of the Sung. Thus:

Table 5. Use of the relation in Sung map-making

	Map squares	Difference Δ in latitude	*li* from map	Difference in *li* = △ × 351
Khai-fêng to Kuang-chou	39	11½°	3900	4036
Khai-fêng to Wu-chhang	13–14	4+	1350	1400+
Khai-fêng to Peking	15+	5−	1500+	1755+

[48] J. B. B. D'Anville, "Mémoire sur le *li*, mésure itinéraire des Chinois" in *Mém. de Litt. tirés des Registres de l'Acad. Roy. des Inscr. et Belles Lettres*; **28**, p. 487, 1761.

[49] Such as the +9th century *Yuan-Ho Chün Hsien Chih*.

[50] The text, as we have seen, continues with the statement that the distance from Lin-i to Thieh-lo is the same, viz. 17°4. This suggests that the latitude difference of both places (17°3) was confused with the latitude (17°4) of Lin-i, because $6112 = 17{\cdot}4 \times 351{\cdot}267$, while $17{\cdot}3 \times 351{\cdot}267 = 6077$. The distance to the Thieh-lo Turks is an independent estimate coming from the Chên-Kuan reign-period (+627 to +649).

[51] Cf. *SCC*, vol. 3, p. 547.

Short distances were no doubt taken from the same source as those given in the gazetteers (cf. *Thung Tien* and *Yuan-Ho Chün Hsien Chih*) and measured in short Thang *li*. Thus:

Chhang-an to Lo-yang = 835 *li* = 8+squares
Chhang-an to Khai-fêng = 1280 *li* = 13 squares

The meridian distance Khai-fêng to Peking was probably also estimated from ground measurements. At 250 *li* per degree (i.e. corresponding to the normal short Thang *li*), the distance should be just under $12\frac{1}{2}$ squares, but distances not due N–S and E–W were calculated by adding together intermediate distances, and tend to be exaggerated (e.g. a stretching out of the Yangtze River, Shantung province, etc.). There is no proof that the distances between Khai-fêng and Wu-chhang, and between Khai-fêng and Kuang-chou were not obtained by the same kind of additive process, but the close correspondence to $1^\circ = 351$ *li* is suggestive. They would probably have been even more exaggerated otherwise.

7. Reduction of Shadow Observations

We next proceed to analyse I-Hsing's recorded astronomical data, in order to find out whether they were all purely observational, or whether some were deduced from the observations by calculation. Furthermore, we shall try to obtain an estimate of the internal accuracy of the observations and calculations.

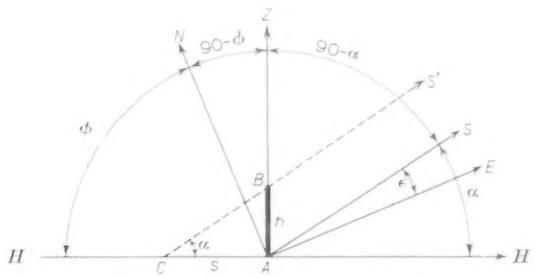


Fig. 5. Gnomon shadows

A = observer
HH' = his horizon
Z = his zenith
E = equator,
N = North Pole
ϵ = obliquity of the ecliptic,
$HAS = ACS' = \alpha$ = direction to Sun at Summer Solstice,
$AB = h$ = height of gnomon,
$AC = s$ = length of shadow at Summer Solstice.

From the figure follows:

$$\tan \alpha = h/s \qquad \ldots . (1)$$

and since at Summer Solstice:

$$\angle NAE = \angle NAZ + \angle ZAS + \angle SAE = 90^\circ,$$

we obtain the relation between α_S (i.e. the Sun's altitude at Summer Solstice) and the geographic latitude ϕ of the station:

$$\phi = (90 + \epsilon) - \alpha_S. \qquad \ldots . (2)$$

Similarly, at Winter Solstice:

$$\phi = (90 - \epsilon) - \alpha_W \qquad \ldots . (3)$$

and at the Equinoxes:

$$\phi = 90 - \alpha_E, \qquad \ldots . (4)$$

where $\tan \alpha_W = h/s_W$ and $\tan \alpha_E = h/s_E$; s_W and s_E being the shadow lengths at Winter Solstice and at the Equinoxes, respectively.

The angles α and ϕ, which were derived from the ratios h/s, are of course in European degrees and independent of the fact that h and s were originally both expressed in Chinese feet.[52]

First let us examine the shadow lengths for internal consistency. Using equations (2), (3) and (4), we can eliminate ϕ and thus find the following relations between the α-values:

$$\alpha_S + \alpha_W = 2\alpha_E \qquad \ldots . (5)$$

$$\alpha_S - \alpha_E = \epsilon \qquad \ldots . (6)$$

$$\alpha_S - \alpha_W = 2\epsilon \qquad \ldots . (7)$$

The following Table 6 gives the various α-values based on Table 1, and also in the sixth, seventh and eighth columns three quantities derived from Equations (5), (6) and (7), respectively.

Table 6. Solar altitudes

Serial No.	Station	α_S	α_W	α_E	$\frac{\alpha_S+\alpha_W}{2} - \alpha_E$	$\frac{\alpha_S-\alpha_W}{2}$	$\alpha_S - \alpha_E$
1	Thieh-lo	62°42′	15°18′	39° 2′	− 0°2′	23°42′·0	23°40′
2	Wei-chou	74 2	26 43	50 21	+ 0 1′·5	23 39·5	23 41
3	Hua-chou	78 54	31 36	55 12	+ 0 3	23 39	23 42
4	Pien-chou	79 10	31 55	55 30	+ 0 2	23 37·5	23 40
5	Yang-chhêng	79 27	32 11	55 47	− 0 2	23 38·5	23 40
6	Hsü-chou	79 48	32 33	56 8	+ 0 2	23 37·5	23 40
7	Yü-chou	80 19	32 52	56 34	+ 0 2	23 43·5	23 45
8	Lang-chou	84 30	37 14	60 48	+ 0 4	23 38	23 42
9	Chiao-chou	92 22	45 13	69 53(?)	(− 1 5′·5)	23 34·5	(22 29)
10	Lin-i	94 5(?)	49 14	71 40	− 0 0′·5	(22 25·5)	(22 25)

The sixth column of Table 6 is the most interesting one, since it relates only the quantities originally given in the text. According to Equation (5) we should have

$$\frac{\alpha_S + \alpha_W}{2} - \alpha_E = 0.$$

Indeed, this relation is very nearly fulfilled; the largest deviation (excluding the aberrant case of station 9) is 4′ and the mean is $1\frac{1}{2}$′.

The accuracy achieved is uncanny. It can easily be seen that the computed shadow lengths which are quoted to 3 significant figures are indeed accurate in the last quoted figure, that is, to 0·1 inch or nearly 1 part in a thousand.

In the first place, this clearly excludes the possibility that all the shadow lengths represent observed quantities. The observational methods used would simply not be capable of yielding such an accuracy (cf. our discussion on p. 22 below of the penumbra at the end of the shadow). Thus just as the longer terrestrial distances

[52] For the two most southerly stations 9 and 10, where $s_9 = -0{\cdot}33$ and $s_{10} = -0{\cdot}91$, note that the α_S values exceed 90°.

were computed, so also were the Winter Solstice and most of the equinoctial shadow lengths.

But secondly, no geometrical construction of such an accuracy would be conceivable either. To enable divisions of this smallness, i.e. about 2′ of arc, to be read off with any certainty they would have to be separated by, say, 1 cm. This magnification would have required a circle of radius exceeding 17 m., or a room at least the size of a large palace hall. Such halls were of course available in China in those days, but that does not make it more probable that a perfectly plane surface and a reading device of the required perfection were really technical possibilities. So prominent a piece of apparatus would surely have been mentioned in our texts, but they are silent on it, nor have we ever encountered any statement in relevant Chinese astronomical writings which could be interpreted as a reference to it.

There can thus be little doubt that these shadow lengths represent the results of computations starting from one set which alone was observational. Such computations would have required tables of tangents, or their equivalent in sine tables; but these would have had to be accurate to 1 part in 500, and intervals of the order of 5′ of arc (a tenth of a *tu*) would have been needed. That such an accuracy was available in the +8th century is a somewhat unexpected conclusion, and we feel that the point merits particular note. Our texts certainly give clear indications of the use of diagrams in aiding the performance of the calculations, but perhaps it might be unwise to interpret the word *thu* too rigidly or exclusively in geometrical terms, for tables may perhaps also be implied by it.

Is there now any other evidence for the possession by Chinese astronomers of this period of even the beginnings of trigonometrical tables? It is of great interest that this has already been suggested as a strong probability on entirely different grounds. Greek astronomy did not know the trigonometric functions as such, but tables of chords were drawn up, first by Hipparchus (ca. −130), and then in more refined form by Ptolemy (ca. +140).[53] Nevertheless the basic steps in spherical trigonometry were taken by Menelaus of Alexandria (ca. +100), who first enunciated the "rule of six quantities". But it was the Indians who brought trigonometry into its modern form. The notion of sines and versed sines appears for the first time in the *Pauliśa Siddhānta* shortly after +400.[54] Āryabhaṭa (ca. +510) was the first to give a special name to the function, and to draw up a table of sines for each degree. His contemporary Varāha-Mihira, in the *Pañca Siddhāntikā* (ca. +505; now our only source for the original text of the celebrated *Sūrya Siddhānta*), gave formulae which in modern terms would comprise both sines and cosines. Then on the one hand the Indian work was taken over by the Arabs and transmitted to Europe,[55] while in the other direction Indian monks or lay mathematicians who took service with the Chinese Imperial Bureau of Astronomy spread the new branch of mathematics further East.[56]

All through the +7th and +8th centuries three clans of Indian experts were resident at the Chinese capital. Their names were Chhüthan (Gautama), Chiayeh (Kāśyapa), and Chümolo (Kumāra). By far their greatest representative was the astronomer and mathematician Chhüthan Hsi-Ta, at the height of his powers just

---

[53] Cf. e.g. D. E. Smith, *History of Mathematics*, Ginn, Boston, 1925, vol. 2, pp. 604 ff.; B. L. van der Waerden, *Science Awakening*, Groningen, 1954, pp. 206 ff., 271 ff.
[54] Cf. G. Sarton, *Introduction to the History of Science*, vol. 1, p. 387.
[55] Cf. A. Mieli, *La Science Arabe*, pp. 83, 108, and a summary in *SCC*, vol. 3, p. 108.
[56] For further details concerning what follows see *SCC*, vol. 3, pp. 202 ff.

when I-Hsing and Nankung Yüeh were busied with their meridian arc. In +729 he finished the *Khai-Yuan Chan Ching* (Treatise of the Khai-Yuan reign-period on Astronomy and Astrology), a work which still today constitutes our greatest surviving thesaurus of ancient Chinese astronomical quotations and fragments. Some parts of this had been ready, however, as early as +718, notably chapter 104 which consists of a translation into Chinese under the name *Chiu Chih* of the +6th-century *Navagrāha* ("Nine Upholders") calendar system of Varāha-Mihira. This was the first occasion on which a zero symbol appeared in a Chinese text (ancient Chinese mathematical notation having always used a blank space), but an innovation more immediately relevant is the table of sines which the chapter contains. Yabuuchi, who has recently called attention to this remarkable fact, reproduces the table in his paper.[57] It is typically Indian in that it tabulates at intervals of 3°45′, a value derived by continuous bisection of the angle of 60°, the cosine of which was known to be $\frac{1}{2}$. Very probably this was by no means the first of such tables to be known in China, since books with such titles as *Po-lo-mên Thien-Wên Ching* (Manual of Brahmin Astronomy) had been circulating at least as early as +600, as we know from the bibliography of the official history of the Sui dynasty.

The upshot of all this is that it now seems very plausible that I-Hsing and Nankung Yüeh had at their disposal trigonometrical tables of sufficient accuracy to enable them to compute the sets of figures which we have been examining. The fact that they made such use of them may indicate a natural predilection for methods which were new, precise, and interesting.[58] We shall return later on to the question of how they could seemingly so fully substitute for a body of practical observations.

Another interesting result can be deduced from our data, namely the value of the obliquity of the ecliptic used in I-Hsing's calculations. This value follows from Equations (6) and (7), and the individual values so found are listed in Table 6, columns 7 and 8. There the overall mean is, for stations 1–8:

$$\epsilon = 23°40'.$$

The value of ϵ extrapolated from modern data would be 23°36′, but all the individual values of our set exceed this. It is of much interest that the mean found is very nearly equivalent to 24 *tu*, and indeed the *Chiu Thang Shu*, in a passage (p. 5b) just preceding the account of the measurement of the meridian arc, states that this value was considered correct in I-Hsing's time. In fact it agrees exactly with the value given by Li Shun-Fêng, his great predecessor, in the previous century.[59] It was thus the figure generally accepted in the early Thang period.

The shadow measurements at stations 2 and 8 in Table 1 deserve special remark because they are the only ones for which some equinoctial shadow lengths among those preserved may be actual observations. Hence we may have a clue as to the accuracy of the measurements. In Table 6 the ES values used to derive α_E were taken from the commentary in the *Chiu Thang Shu*; if the ES values from the text

[57] Yabuuchi Kiyoshi, "Indian and Arabic Astronomy in China"; art. in *Silver Jubilee Volume of the Zinbun Kagaku Kenkyūso* p. 585. Kyoto University, Kyoto, 1954.

[58] Thus it is already clear that some modification will be needed in the opinion expressed in *SCC*, vol. 3, p. 203, commenting on the remarkably little influence exercised by the Indian scientific men on the course of Chinese astronomy. "The equatorial mansions remained as before, the circle continued to have 365¼°, Indian trigonometry was not taken up, the zero symbol slumbered for another four centuries, and naturally enough the Greek zodiac remained buried in bizarre transliterations." Nevertheless the statement is still broadly true.

[59] See *SCC*, vol. 3, p. 289. Li Shun-Fêng lived from +620 to +680.

itself had been chosen there would have been a difference of 0·5 and 0°8, respectively. This is of the same order of magnitude as the mean difference between calculated latitude and modern geographical latitude, i.e. 0°33 (see p. 24 below).

8. Reduction of Latitude Data

We now turn to a discussion of the latitudes of the observing stations derived from the data in Table 6.

Table 7. Polar altitudes

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	Thieh-lo	52	51·69	52·76 ±	(+0·31)	(−1·07)	52 ±	114 ±
2	Wei-chou	40	40·19	40·38	−0·19	−0·19	39·8	114·5
2a	Thai-yuan	—	37·42	38·22	—	(−0·80)	37·70	112·5
3	Hua-chou	35·3	35·26	36·07	+0·04	−0·81	35·55	114·7
4	Pien-chou	34·8	34·98	35·31	−0·18	−0·18	34·80	114·55
5	Yang-chhêng	34·7	34·70	34·93	0·00	−0·23	34·43	113·03
6	Hsü-chou	34·3	34·34	34·65	−0·04	−0·31	34·15	114·38
7	Yü-chou	33·8	33·82	34·09	−0·02	−0·27	33·60	114·35
7a	Hsiang-chou	—	31·17	32·46	—	(1·29)	32·01	112·1
8	Lang-chou	29·5	29·58	29·42	−0·08	+0·16	29·0	111·60

Table 7 gives:

(1) serial number of station.
(2) name of place.
(3) latitude of station, from Table 1 (Polar altitude in Chinese degrees (*tu*)).
(4) the same, computed from shadow length at Summer Solstice in Chinese degrees.
(5) the same, taken from Ting Wên-Chiang's atlas (1934), in Chinese degrees.
(6) difference: (3) minus (4).
(7) difference: (4) minus (5).
(8) latitude from Ting Wên-Chiang's atlas (1934), in European degrees (°N).
(9) longitude of station from same atlas, in European degrees (°E).

As to the location and correct identification of the several stations we should note:

Station 1: The location in the vicinity of Lake Baikal is discussed by Tung Tso-Pin.[60]

Station 2: The position of Wei-chou (Hêng-yeh chün) is described on p. 36 of Tung Tso-Pin's book as being north-east of the present Wei-chou district in Chahar province, 39°8 N and 114°5E.

Station 3: The positions of the stations Nos. 3, 4, 6 and 7 are of particular importance, since this central chain forms the basis of our subsequent conclusions. The positions which were given the tabulated names in the +8th century are not necessarily the same as those so named today. It appears however that with the exception of No. 3 they have changed by one or two kilometres only. Station 3 was situated in the +8th century on the southern bank of the Yellow River; but as a result of the river changing its course, modern Hua-chou is well North of it. The *Encyclopaedia of Chinese Geographical Names* states that old Hua-chou was about 20 *li* East of modern Hua-chou, the longitude of which is 114°6. Since we can assume that the term *li* in modern Chinese means (unless qualified by *kung*) half-kilometres, we thus obtain for the value of the correct longitude 114°71.

Station 5: This is the ancient Central Astronomical Observatory. According to

[60] Tung Tso-Pin *et al.*, p. 48.

Tung Tso-Pin, p. 28, its position was 34°26′ N and 113°2′ E; another modern author, Yabuuchi,[61] gives 34°63 and 113°00 E.

Station 9: Situated in Indo-China; its longitude cannot be greater than 106° E.

Station 10: In Indo-China; longitude, say, 106° ± 1°.

We now form from Table 7 the mean of column (6), i.e. the mean difference between the recorded polar altitude and the corresponding value derived from the shadow at Summer Solstice. We obtain −0°07, with sign considered, and 0°08 with sign not considered. Evidently this is of the same order of magnitude as found in Table 6 for the inter-agreement of the three different types of shadow lengths given in the texts. Hence we conclude that the recorded polar altitudes, like the longer terrestrial distances, and most of the shadow lengths, must also have been computed. We note, of course, that for stations 1 and 2 the discrepancies are rather large, but in these cases the Polar Altitudes are only given rounded off to the nearest degree; this leaves only another large deviation, that of station 4.

Turning now to consider the geographical locations of the observing sites, as recorded on modern maps, we form the mean values of column (7) in Table 7 for the main stations 2, 3, 4, 5, 6, 7, and 8, i.e. the mean difference between the latitude as calculated from Summer Solstice shadow length and the latitude on the map. The result is −0°31, sign considered; and 0°33, sign not considered. This result appears to show that the actual observing sites were on the average rather South of the geographically identified places (in 6 cases out of 7). At a mean latitude of 35° N a deviation of 0°31 corresponds to 28 km.

Nevertheless there may surely be further significance in the fact that the deviations have nearly all the same sign. This may point to a systematic error in the method of observation caused by an incorrect convention used to define the end of the shadow. For the end of any shadow is hazy due to the finite diameter of the Sun, which causes the intensity to decrease not sharply but gradually. In order that the formulae we have been using should be correct, it would be necessary to define the end of the shadow as the mid-point of this penumbra (that is, the indistinct hazy end of the shadow), so that the observations really refer to the centre of the Sun's disc. If, on the other hand, I-Hsing had observed the end of the blackest part of the shadow, he would have got a systematic error of −0°25 (radius of the Sun) in his latitudes.

Even this leaves us with an unexplained −0°06, and although small the deviation is still systematic. However, the appearance of the boundary would depend very much on the geometry of the top of the gnomon and the state of the atmosphere, as well as on the subjective judgment of the observer. It appears that the observers missed the outermost faintest extension of the shadow and measured a point somewhere further inside, thus producing the spurious South deviation. Here we may keep in mind that 0°06 corresponds to less than 0·1 inch.

Five hundred years later these difficulties were well appreciated, and Kuo Shou-Ching in +1279 developed a special pin-hole device, the "shadow-definer", for focussing the image of the cross-bar at the top of his 40 ft. gnomon. Laplace considered that these thirteenth-century observations were among the most accurate ever made of solstitial sun-shadows.[62] Kuo Shou-Ching also caused a series of polar altitude observations to be made at this time, much more extensive than those of

[61] Yabuuchi Kiyoshi in *Zuitō Rekihoshi no Kenkyū* (Studies on Sui and Thang Calendrical Science); in Japanese.

[62] See *SCC*, vol. 3, p. 299.

I-Hsing's survey. The data are recorded in the *Yuan Shih* (History of the Yuan Dynasty), ch. 48, pp. 12b ff., and it would be interesting to subject them to a similar study.

A question may be raised as to the reliability of the set of figures we have been analysing. While one must naturally avoid casting any doubt on the transmission of the text if possible, one cannot but be struck by the fact that in Table 7 the data of stations 3, 4 and 5 in cols. 3 and 4 would agree extremely well with those in col. 5 for stations 4, 5, and 6. One cannot but ask the question therefore whether it is not possible that the columns of figures in the archives of the Bureau of Astronomy and Calendar could have been misread by the writers of the Bureau of Historiography during the compilation of the Thang official histories. Admittedly this would leave us without data for station 3 (Hua-chou) and with a set of figures for some station intermediate between Nos. 6 and 7. At any rate, similar instances of textual confusion are by no means unknown.[63] But the fact that the texts as we have them date from such a short time after the work of I-Hsing (see p. 7 above) may be considered to render such a confusion less likely in this case.

Here we must recur to station 5. As we saw above, Yang-chhêng was for centuries the seat of the central observatory of China. This was certainly a station served by experienced astronomers where we should expect good observations. Indeed column (7) of Table 7 shows that the observed latitude agrees well with the geographic position on modern maps. The difference is $-0°.23$, which is the third smallest deviation of all. Reducing this by $+0°.25$ for the Sun's radius (see above), the agreement becomes perfect ($+0°.02$). We may also note that there exist two complete sets of observing data for this station (see notes to Table 1).

Of course our main attention concentrates itself on the central chain (stations 3, 4, 6, 7) of I-Hsing's observing sites. A closer examination of the data of Table 3 (page 15) and the notes to Table 7 shows that according to the assumed identifications I-Hsing did not get his meridian arc quite straight, for stations 3 and 4 are off the meridian defined by the sites 6 and 7. Consequently the differences in polar altitude, which correspond to distances along the meridian, should actually be corrected to correspond with the actual distance as measured along the ground. Consider, for example, station 3, which gives the largest deviation in Table 7. At its latitude of $34°.8$ a longitude difference of $0°.1$ corresponds to $0°.082$ in latitude. According to p. 21 the East deviation of Hua-chou amounts to $0°.31$ in longitude, which corresponds to $0°.25$ for the distance from a point where the perpendicular through the station to the meridian intersects the latter. We thus have a right-angled triangle, the other side of which is the angular distance from this point to station 4, namely, $0°.28$. The hypotenuse is the "equivalent latitude difference" which corresponds to I-Hsing's measured distance, 198·60 *li*. Pythagoras' theorem gives for it $0°.38$. We mention these figures only as an illustration, for we consider them too tentative to be applied to I-Hsing's original data as summarized in Table 3. The value of 351 *li*/degree, derived from the three central pairs, agrees with the corresponding rounded-off values of the latitudes as derived from the Summer Solstice shadows within $0°.1$.

---

[63] Cf. the mistakes in the comet registers in Ma Tuan-Lin's *Wên Hsien Thung Khao* (Ho Ping-Yü, *Vistas in Astronomy*, ed. A. Beer, vol. 5, p. 117, Pergamon Press, London–New York, 1961), and the mediaeval garbling of the *Mo Ching* (*SCC*, vol. 2, p. 171).

There have been some suggestions that I-Hsing did not use the normal short Thang *li*, but some specially defined *li*, as a basis of his astronomical and geodetic system. This suggestion has arisen because his final result of 351 *li*/degree is so different from that of the normal short Thang *li* (250 *li*/degree, see Table 2). Now that we can judge to some extent the quality of I-Hsing's observations, this discrepancy is no longer serious. It would be just within the bounds of probable error that the entire discrepancy lies in the latitudes derived for stations 3 and 7. Judging alone from the difference between his observed values and modern geographical latitudes, for which we may take a mean of 0°33 for one station (page 22), the total measured arc 3–7 might have extended to as much as 2°3 (instead of 1°5, cf. Table 3) leading to 230 *li*/degree, which is within the range of the short Thang *li*.

Of course we have no means of judging the accuracy of the measured distances, and in particular we recall that the line measured was not exactly a meridian line, the total East–West deviation of stations 3 and 7 being of the order of 40 km. In view of what has been said above, we do not propose to pursue this discussion further or to stress the fact that the normal short Thang *li* seems to have been the unit used by I-Hsing. Without the preceding discussion of the margin of error, one might have felt inclined to see more relevance in suggestions about other possible units.

For example we may refer to an edict of +721, according to which a "small measure" was to be used for astronomy, medicine, and imperial paraphernalia, while the "long measure" was to be applied to everything else, whether official or private.[64] I-Hsing could have obtained an approximately 300-metre *li* by taking the ancient Chou foot of 0·198 metre. It is not at all unlikely that Chou foot-rules were known in his time, and such a reversion to the ways of "the former kings" would have been quite in the Chinese taste. Furthermore, Wu Chhêng-Lo in his *History of Chinese Weights and Measures* records[65] that there was a *pu* (pace) of only 5 ft during and after the Thang Dynasty (+618 to +907). Generally there were 360 of these *pu* to the *li*, but we know that in I-Hsing's system there were 300 *pu*.[66] This was the old Chou system. Indeed, if I-Hsing's *li*/degree ratio be expressed in modern terms, 1° would be equal to 356·2 *li*, and the value of his foot would be 0·174 m. at 6 *chhih* (feet) to the *pu* (pace), and 0·206 m. at 5 *chhih* (feet) to the *pu* (pace). Yet the view of Tung Tso-Pin who says (*loc. cit.*, p. 43) that Nanking Yüeh and I-Hsing intended to use the Khai-Yuan foot (0·2456 m.), seems more probable.

9. Discussion

The question arises of what significance to assign to the value of 351 *li* 80 *pu* given to the degree in the Thang texts. There can be no doubt that this was derived only from the polar altitudes calculated for the four central stations (3, 4, 6 and 7). We have seen that the distances of the outer stations were computed from this value, and that the latter seems to have been used by later cartographers. But although this figure appears in the texts as the principal result of the whole investigation, it was bound to be very inaccurate, being based on such a short meridian line of some two hundred kilometres, with a correspondingly large percentage error.

---

[64] Wu Chhêng-Lo, *Chung-Kuo Tu-Liang-Hêng Shih* (History of Chinese Weights and Measures), 1st ed. p. 387, 2nd. ed. p. 253.

[65] 1st ed p. 96, 2nd ed. p. 61.

[66] That this was so is clearly implicit in his figures for the distances (see Fig. 4): (3–4)+(4–6)+(6–7) =526/270; therefore 1 *li* = 300 *pu*.

In later times better approximations were reached. Gaubil says[67] that in the Sung (+1001) an arc of 3° latitude was taken as equivalent to 1000 *li*, i.e. 1° = 333 *li*.[68] This was a little nearer to the correct value of 250 *li* (see p. 15 above).

Did I-Hsing attempt to derive from his measurements a value for the circumference of a spherical Earth? It is impossible to say.

The stated motivation of I-Hsing's expedition was to test the old belief that the Sun shadow lengths changed 1 inch for a distance of 1000 *li* on the ground in the North-South direction. This had been sharply called in question by the observations in Lin-i and An-nan in the +5th century. The result of the field survey was to prove the falsity of the old belief, and I-Hsing and his collaborators established the true relation between distance and change in polar altitude. In this his knowledge of Indian and even Hellenistic astronomical achievements may possibly have helped him, though nothing in his proceedings was foreign to Chinese astronomy.

Although no record remains in the texts that I-Hsing made any calculations to obtain from his data the dimensions of a spherical Earth, certain Chinese cosmological schools had from antiquity onwards assumed its sphericity.[69] This would have been very well known to him, even if his contemporary Chhüthan Hsi-Ta had not at the same time been engaged in compiling a great collection of ancient and early mediaeval astronomical writings. Moreover, his acquaintance with Indian and Hellenistic astronomy, gained as an outstanding Buddhist scholar, may well have informed him of the previous estimates of the Earth's circumference. There is thus no reason why I-Hsing should have hesitated to use in this way the data which his observers collected. Besides, it is hard to see how he could have given a constant *li* per degree, if he had not had at least some *arrière-pensée* of a curved Earth's surface.

Yet it is of much interest that in the *Chiu Thang Shu*, after the conclusion of the account of the survey itself, there follows (pp. 8b ff.) a cosmological discussion which attempts to reconcile the results obtained with the hypothesis of a flat Earth. Using I-Hsing's figures, calculations of the size of the universe on the basis of some of the ancient cosmological ideas, and in rather archaic style, are set forth. In the end the writer is reduced to saying: "But looking at it in this way, of what value for the cultivation of human morality are the hair-splitting discussions of Wang Chhung and Ko Hung?"[70]

Thus if I-Hsing himself had any conception of a spherical Earth, he must have kept it to himself and his immediate disciples such as Ta-Hsiang and Yuan-Thai. There is little ground for supposing that such a notion would have been suppressed on doctrinaire grounds as unorthodox by the Confucian scholars of the time. Confucian orthodoxy was in fact particularly weak just then. The fact that I-Hsing's results discredited the beliefs of the "scholars of former times" (*hsien ju*) was in no way glossed over; on the contrary it was rather emphasized. After all, *ju* was the only

[67] *Loc. cit.* (footnote 41) p. 97.

[68] Hence d'Anville's supposition (see footnote 48) that 351 *li* must be a mistake of Gaubil's for 331 *li*; but this is not so as all the texts read 351.

[69] See *SCC*, vol. 3, pp. 216 ff., 498 ff.

[70] P. 9b. These were ancient cosmologists living in the +1st and +4th centuries respectively.

One wonders whether we do not have to deal with differences of opinion between the Astronomer-Royal Nankung Yüeh and the imperially favoured Buddhist monk-astronomer I-Hsing. In the year after the latter's death, +728, his calendar system, the *Ta Yen Li*, was published, but in +733 Chhüthan Chuan (another of the Chhüthan clan) attacked it, declaring that it was only a plagiarism of the *Navagrāha* calendar which Chhüthan Hsi-Ta had translated in +718. In this action Chhüthan Chuan was supported by Nankung Yüeh.

word current for Confucians, even if often used with wider implications. Moreover the *Chou Li* (Record of Institutions of the Chou Dynasty), the book which enshrined the discredited doctrine, was still fully canonical, and serious controversy about its authenticity did not arise until much later. In fact one must say that all the scholars concerned in I-Hsing's time showed a relatively enlightened conception that age-old beliefs about the universe must bow to improved scientific observations.

10. Appraisal

Coming now to the conclusion of our study, we recall that the analysis of the set of figures recorded in the Thang texts brought to light as it went on a series of rather unexpected findings. First it could be shown that the ground distances between the more distant stations were not measured but assessed by extrapolation on the basis of the result for the short central line of stations. Secondly, it appeared that all the Winter Solstice shadows and the great majority of the equinoctial shadows recorded were not measurements but values calculated from the Summer Solstice shadow series. Thirdly, it could be demonstrated that the recorded polar altitudes themselves were also not measurements but values calculated from the Summer Solstice shadow data. Other considerations in their turn led to the conclusion that adequate graphic methods for the performance of these computations would have been very improbable, and hence that tables of trigonometrical functions must have been available to the computers. Nothing in this offends historical probability.

From these remarkable facts it would be easy to conclude that in the course of the survey very few actual observations of any kind were made, and that theory took an overwhelming precedence over practice in +8th-century China. But such a judgment would be superficial, and indeed misleading. Our texts expressly indicate that the expeditions were real and thorough. In +724 observations by "Commissioners for Shadow Measurement" were ordered at Chiao-chou (in modern Indo-China) and at least eleven other places, covering a total arc of some 2500 km. In the following year the central line of stations was set up, and measurements collected. Only after two or three years did the Commissioners return to the capital and confer with I-Hsing upon their multifarious results.[71]

It certainly seems quite strange at first sight that the final set of figures presented by I-Hsing to the imperial court, and hence destined for transmission to us, should have been mainly computed and not observational. But surely what this means is that at that time it seemed more elegant to offer up a set of "ideal values" computed by the most up-to-date mathematical methods using sines and cosines. Today our minds are so accustomed to statistical methods that it is hard for us to imagine a time when they were not understood. In all probability I-Hsing thought it very undesirable to admit to his final tabulation a mass of raw data showing considerable scatter, and not being able to assess it statistically, he used it only to satisfy himself that his calculated values came about where they should—in fact he probably believed that they were much more reliable than most of the observations. In other words he "constructed a curve to fit". The reason why the Summer Solstitial shadow figures were taken as the primary data, now the only observational portion of the set, was perhaps because I-Hsing considered them the most accurate. In sum, there is no reason for doubting that a quantity of actual measurements of shadow lengths by

[71] The period of their labours is circumscribed by the fact that I-Hsing died in +727.

gnomon scales and most probably also of polar altitudes by armillary sphere quadrants came to rest in the archives of the Bureau of Astronomy and Calendar. We should be glad to have them today.

In whatever way one considers the situation, I-Hsing's work takes an outstanding place in the pre-history of the metric system. It occupies a unique position among the scientific achievements of the early Middle Ages, and it is surprising that in spite of its recognition by some 18th century authors such as Gaubil and d'Anville, it has remained one of the least known.

11. Epilogue

A thousand years later, in the last years of the 17th century, the relations between East and West had long been greatly intensified, and the science of the Renaissance had become well known in China because of the activities of the Jesuit mission. Indeed, it was a distinguished member of the missionary clergy who took up a suggestion of the Khang-Hsi Emperor which anticipated the French Revolution by nearly a century in relating terrestrial measurements of distance to cosmic standards. He was Antoine Thomas (1644–1709), a Belgian from Namur, who arrived in China in 1685, and became Vice-Director of the Bureau of Astronomy. There he began the planning of the great cartographic survey of the Empire,[72] which was finished in +1717; and one of the pre-requisites for this work was the decision to determine the exact length of the *li*. Since estimates varied from less than 200 to more than 250 *li* to the degree, the idea grew up between Khang-Hsi and An To (Thomas) of pinning the *li* specifically to the degree of terrestrial latitude, i.e. to the Earth's circumference.

The final decisions were made in December 1698, and the actual meridian-line work began in 1702.[73] The observations were carried out with remarkable accuracy (gnomons having been erected at both ends of a line in the plains near Pa-shih, south of Peking), and it was ascertained that 1° of terrestrial latitude corresponded to 195 *li* 6 *pu* of the standard in use at the time. This was a 5 ft-bar of iron, kept in the Imperial Palace, representing the "geometrical pace". As the Khang-Hsi Emperor wished to adjust the *li* to the degree in round numbers, it was decided to adopt the figure of 200 *li*, thus involving a diminution of the standard foot by a factor of 40/39. There were thus 72,000 standard paces to the degree, 1200 to the minute, and 20 to the second of arc. At 0·555 km, this *li* was closely similar to the *li* used in our own time.[74]

Thus the *li* was fixed astronomically some ninety years before the metre. For it was not until 1791 that a committee of the French Academy made their famous report in which it was suggested that the metre should be taken as one ten-millionth part of one quarter of the Earth's circumference at sea level.[75]

---

[72] Cf. *SCC*, vol. 3, p. 585.

[73] For the full description and translation by H. Bosmans see *SCC*, vol. 4, pt-1. See also references in J. Needham's article on Chinese observational astronomy in *Vistas in Astronomy*, ed. A. Beer, vol. 1, pp. 67–83, Pergamon Press, London & New York, 1955. Also J. Needham, *Chinese Astronomy and the Jesuit Mission; an Encounter of Cultures*, China Society, London, 1958.

[74] It will be seen that this *li* was nearly equivalent to the long Thang *li* (Table III, above). D'Anville (see footnote 48) adopted 195 *li* to the degree in his map of Asia.

[75] It should be pointed out in this context that there arises, as usual, the question of transmission. The Jesuit work was, of course, well known in France, where it was popularized in writings such as those of d'Anville. But both he and Gaubil misunderstood the work of Thomas and supposed that the Khang-Hsi emperor had hit upon 200 *li* arbitrarily without the intention of making an astronomical definition.

Once again conditions in China had permitted large-scale scientific observation and rational public action based thereon some time before Europe was ready to follow suit.

VOL
4
196

Artificial Satellites

Sir Harrie S. W. Massey
University College, London

1. Introduction*

The Importance of Artificial Satellites

A great deal of information has been obtained about the structure and properties of the atmosphere, solar-terrestrial relations, cosmic rays and meteor phenomena from observations carried out with ground-based equipment or with equipment carried in balloons to altitudes below 100,000 ft. Nevethelesss, there are some important aspects which cannot be investigated in this way and to study them it is necessary to turn to the considerably more expensive technique of using high altitude sounding rockets, capable of attaining altitudes of 100 km or more. It is only in this way that information can be obtained about the spectral intensity distribution of the solar radiation incident on the outer atmosphere. Without this knowledge it is not possible to understand in detail how the absorption of sunlight produces the ionosphere, the ozone layer, etc.

Although the constitution of the solar radiation is the most important matter which requires the use of sounding rockets, there are many others which are either very difficult or impossible to study from ground. These include the variation of electron concentration with height above the F_2 maximum and in the regions above the E and F_1 maxima when it is decreasing with increase of height, the composition of the positive and negative ions at different heights, the height variation of the intensities of the main lines and bands in the night airglow, the fraction of oxygen in the atomic form at different heights, the mass distribution and flux of meteorites, and so on.

Furthermore, study of radiations which are absorbed by the atmosphere need not be confined to radiations from the Sun. Cosmic sources of various kinds may be investigated. A start in rocket astronomy has already been made by Kupperian, Friedman and their associates (Kupperian 1958a, b, c.) (see p. 69).

Important as the sounding rocket is for this work its use is attended by certain serious limitations. Geophysical and solar phenomena vary markedly and often irregularly with time while atmospheric effects often exhibit a strong dependence on geographical location. A sounding rocket completes its trajectory in a matter of minutes so that it is very unsuitable for the study of these variable phenomena. The possibility of wide range in latitude and longitude can also be ruled out on economic grounds, even though the development of cheap rockets for synoptic use has proceeded quite far. It is to meet the requirement of studying phenomena over the

* No detailed discussion of material published after 1959 January 1, has been included but as far as possible some reference is made to important developments reported up till 1959 November 1.

whole Earth, phenomena that may vary rapidly with time, that the artificial satellite comes into the picture.

If a satellite consisting of a container of sufficient size to accommodate a useful weight of equipment is able to circulate for months, it is possible to observe the relevant geophysical, solar or astrophysical phenomena in a continuous manner, both as regards time and geographical location. This has the important additional advantage of providing opportunity for observation of weak radiation from cosmic sources so that the scope for extra-atmospheric astronomical observation is greatly extended. The requirement of adequate lifetime can be met by launching the satellite into an orbit whose perigee is far enough above the surface for atmospheric resistance to be sufficiently small.

Observational techniques which have been developed for the investigations with sounding rockets can be used effectively in a satellite. This applies particularly to the telemetry—the method of transmitting data to ground in the form of coded radio signals. This of course requires the provision of a source of power within the satellite. As the weight which may be carried up to the chosen orbit is limited, the battery capacity which can be used is also limited. This is a more serious restriction to effective life than the effect of air resistance, but will be overcome when solar batteries are adequately developed. These are capable of converting solar radiant energy into electrical energy.

It is important to realize that, to take advantage of the long period of observation possible from a satellite, the instruments employed must be capable of automatic operation, without maintenance, for periods longer than usually required or achieved.

Even with a satellite which contains no observing instruments it is possible to obtain valuable information from sufficiently precise determination of the satellite's orbit as a function of time. Thus, if the perigee of the initial orbit falls within the atmosphere, it is possible, from observation of the rate of change of the orbit parameters, to determine the air drag and hence the air density at heights near the perigee. On the other hand, an orbit with much greater perigee distance will be affected to an inappreciable extent by the residual atmosphere. These latter conditions are suitable for precision orbit determination, directed towards obtaining new information about the Earth's gravitational field and the improvement of geodetic measurements. It is also possible that satellites circulating in these high orbits will be of importance for navigation.

The problem of tracking a satellite with precision is rendered difficult by the very high speed of motion relative to the Earth—about 18,000 miles/hr. Experience so far gained indicates that this problem is best dealt with in two stages. Predicted passages of the satellite must be provided by methods which, while relatively inaccurate, are particularly useful for "acquiring" the satellite. Once these rough predictions are available precise visual and photographic methods may be employed as it is possible to narrow down very greatly the area of sky which is to be scanned. This procedure is particularly desirable also because of the short time per transit in which optical observation is possible—only a short time at dawn or twilight, the satellite not being self-luminous.

For the first stage it is a great advantage for a radio beacon to be included in the satellite so that "acquirement" by radio reception is possible. Although apparent directions of arrival of such signals may be obtained with quite high accuracy,

corrections due to the effect of the ionosphere on the propagation of the signals from the beacon are difficult to apply and reduce the accuracy of the time-position determination. Sometimes radar may also be used for acquirement and this has the advantage that it requires no operating beacon in the satellite.

If the greatest possible accuracy is aimed at for true position determinations by radio-beacon it is necessary to use a radio frequency as high as possible. On the other hand a useful compromise may be made by using one or more frequencies which, while affected by the ionosphere, are not so seriously influenced as to render the radio tracking ineffective for the final stage predictions. Experience has shown that it is possible to choose frequencies so as to meet the first stage tracking requirement and at the same time to provide valuable data about the ionosphere up to the heights at which a satellite circulates.

An important practical question is that of the temperature range of a satellite as this may have a serious effect on the operation of the enclosed equipment.

Once the launching and telemetry problems are solved, a great range of possible and desirable experiments may be carried out. These include the measurement of the intensity of the solar ultraviolet, X- and corpuscular-radiation, of the primary cosmic radiation, of the magnetic field in the high atmosphere, of the intensity of the radiation from the Earth to outer space, of the flux of micrometeorites, of the radiation from cosmic sources, and so on. Data of this kind will be of inestimable value for the understanding of solar-terrestrial relations and for solar- and astrophysics generally. As launching techniques improve and larger satellites are launched it will be possible to introduce more elaborate equipment such as television cameras, telescopes, etc.

In view of the remarkable features of the second Russian satellite it must be pointed out that opportunity is also afforded for space-physiological studies. The technique for carrying out such work has already been developed to a high degree in connection with sounding rockets, it being possible to telemeter to ground data about pulse and breathing rate, blood pressure etc., of dogs and monkeys during rocket flight. The physiological effect of prolonged weightlessness as well as of long exposure to the heavy components of cosmic radiation may, in principle, be studied by including mammals in satellite cargoes. The problem of recovery remains, of course, a difficult one. It is not appropriate in this review, even if it were within the competence of the writer, to discuss whether the purely scientific interest of space physiology would of itself justify including animals as part of satellite equipment. To those for whom space travel is an end in itself, no further justification is required.

The possibilities of probing the surroundings of the Earth far out into interplanetary space by using unmanned instrument carriers are certainly very great. It is clear that the technical resources for this kind of work, requiring high precision in launching control, are developing so rapidly that much new scientific information will be forthcoming in this way during the next few years.

In this review of the various aspects of research with artificial satellites we shall first discuss rather briefly the problems of launching and then turn to consider the nature of satellite orbits particularly as modified by air resistance and by the non-central character of the Earth's gravitational field. The next section will deal with tracking methods. This will be followed by an examination of the factors which determine the temperature of a satellite. The scientific applications of satellites in geophysics, solar physics and astrophysics will form the subject of the next seven

sections after which there will be a concluding section devoted to the first satellites which have been successfully launched.

2. Launching Problems

The first thing to decide in the planning of a satellite launching programme is the minimum altitude or surface-perigee distance of the initial orbit. In order to have a reasonably long life this altitude must be not much less than 200 km—the first Russian satellite had a life of about 3 months and its minimum altitude was close to 200 km. The accuracy necessary in launching in order to meet this requirement is high.

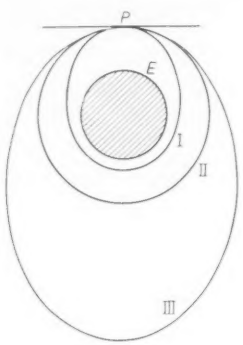


Fig. 1. Illustrating the effect of projection velocity on a satellite orbit.

E = surface of the Earth.

I, II and III are orbits resulting from projection in the horizontal direction at P with launching speeds respectively less than, equal to and greater than the speed V_0 for a circular orbit.

Launching Accuracy

Suppose the satellite is projected with a velocity V at a distance R from the centre of the earth in a direction making the angle θ with the horizontal.

If the Earth's gravitational field is taken to be central, the satellite will describe an elliptic orbit of major axis $2a$ and eccentricity e where

$$\frac{1}{a} = \frac{2}{r} - \frac{V^2}{\mu} \quad , \qquad \ldots . (2.1)$$

$$a(1-e^2) = \frac{R^2V^2\cos^2\theta}{\mu}, \qquad \ldots . (2.2)$$

$$\mu = GM. \qquad \ldots . (2.3)$$

G being the universal constant of gravitation and M the mass of the Earth.

The maximum and minimum distances of the orbit from the Earth's surface, h_a and h_p respectively, are given by

$$h_{a,p} = a(1 \pm e) - R_0, \qquad \ldots . (2.4)$$

where R_0 is the radius of the earth. We therefore have from (2.1) and (2.2)

$$h_{a,p} = \frac{R\mu}{2\mu - RV^2}\left[1 \pm \left\{1 + 2\frac{RV^2\cos^2\theta}{\mu}\left(\frac{V^2R}{2\mu} - 1\right)\right\}^{\frac{1}{2}}\right] - R_0 \qquad \ldots . (2.5)$$

If $V = V_0 = (\mu/R)^{\frac{1}{2}}$ and $\theta = 0$ the orbit will be a circle of radius R.

We consider a situation in which the aim of the launching is to produce a circular orbit of sufficiently large radius. Owing to launching errors the actual orbit will be an ellipse of low eccentricity so we may write

$$V = V_0 + \Delta V, \quad \theta = \Delta\theta.$$

Treating $\Delta V/V_0$ and $2\Delta\theta/\pi$ as small we now find

$$h_{a,p} = \frac{R}{1 - 2\Delta V/V_0}\left[1 \pm \left\{\Delta\theta^2 + 4\left(\frac{\Delta V}{V_0}\right)^2\right\}^{\frac{1}{2}}\right] - R_0 \qquad \ldots . (2.6)$$

If $\Delta^\theta = 0$,

$$h_{a,p} = \frac{1 \pm 2|\Delta V|/V_0}{1 - 2\Delta V/V_0}R - R_0. \qquad \ldots . (2.7)$$

Hence if $\Delta V > 0$ (see Fig. 1), $h_p = R - R_0$ so that the launching height is the minimum height. The maximum distance increases to

$$R - R_0 + 4R\Delta V/V_0. \qquad \ldots . (2.8)$$

On the other hand if $\Delta V < 0$ (see Fig. 1), $h_p = R - R_0 - 4|\Delta V|R/V_0$ and the minimum height is less than the launching height which in fact becomes the maximum distance.

It is clear from these considerations that a deficiency in launching speed is much more serious than an excess as the former alone leads to a decrease in perigee height below the launching height.

If now we suppose $\Delta V = 0$ we have

$$h_{a,p} = (1 \pm |\Delta\theta|)R - R_0. \qquad \ldots . (2.9)$$

An error in launching direction is equally serious in either sense. The effect of such an error on the minimum distance may be reduced by an excess launching speed. It follows that to launch safely from the lowest height compatible with a long satellite life, it is desirable to have available a launching speed as large as possible in excess of the circular value V_0. The only penalty paid for this is an increase in apogee distance and this is unimportant. Fig. 2 illustrates how effective the compensation is between the errors for a particular representative case. In Fig. 2 the error in launching direction which may be tolerated for different available launching speeds, and consequent apogee distances, is given.

Choice of Launching Sequence and Guidance Conditions

The problem of optimizing the launching sequence and guidance conditions so as to meet the requirements of altitude and launching velocity is a complex one and must take account of many factors such as the state of development and availability of components etc. No complete discussion of the subject has been published though various aspects have been described (see for example LAWDEN 1955).

LINES (1957) has discussed the problem of guidance in order to place a satellite in a nearly circular orbit with maximum payload and a minimum expenditure of fuel.

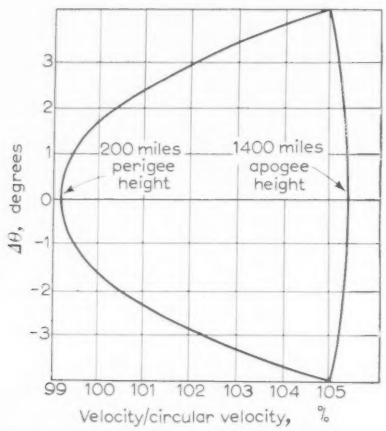


FIG. 2. Illustrating acceptable tolerances in direction and speed of launching for an orbit with perigee height not less than 200 miles and apogee height not greater than 1,400 miles. The circular velocity is that to produce a circular orbit at a height of 300 miles.

If atmospheric effects are neglected it is clear that the sequence which leads to maximum fuel consumption is one in which the vehicle is launched vertically until the desired height is reached and then given the necessary horizontal velocity. This is because all the fuel necessary for the latter stage must be carried to the required height. The optimum trajectory may be obtained as follows.

Consider a projectile fired from the Earth's surface with a velocity V. It will describe an ellipse with the Earth's centre as one focus. The speed at the apogee of the ellipse will reach a maximum when the angle of elevation is zero so that the optimum trajectory for given V is an ellipse tangential to the Earth's surface at the firing point. This is because the minimum amount of fuel will be required to increase the velocity at apogee to the desired extent to produce a nearly circular orbit. The projection velocity V required to reach a height h is given by

$$V^2 = \frac{2gR(R+h)}{2R+h} \qquad \ldots . (2.10)$$

where R is the radius of the Earth. Thus, for $h = 200$ miles, $V = 2{\cdot}63 \times 10^4$ ft/sec. The increase in speed necessary to convert the elliptic trajectory into a circular one at the apogee is given by

$$R\left(\frac{g}{R+h}\right)^{\frac{1}{2}}\left\{1-\left(\frac{2R}{2R+h}\right)^{\frac{1}{2}}\right\} \qquad \ldots . (2.11)$$

and for $h = 200$ miles comes out to be 323 ft/sec.

These considerations must be modified when allowance is made for the effect of the atmosphere. To minimize the amount of energy required to overcome the atmospheric drag it is an advantage to launch the vehicle vertically so as to pass through the denser atmosphere as quickly as possible. Once this has been passed the

preceding considerations apply at least qualitatively. In any case guidance must be supplied to ensure that the various stages in the sequence take place as planned, so that the vehicle trajectory is turned from the vertical at the appropriate time, that the optimum aspect angle is preserved and the different rocket stages come into operation correctly.

It is not possible with chemical fuels to carry through the whole launching operation with a single stage rocket but it may be done by using a number of stages. An advantage is gained because, once the stage is burnt out, it may be jettisoned, so reducing the mass which has to be accelerated in the next stage. An important choice to be made is the number of stages to be used. Fig. 3 illustrates the variation in the

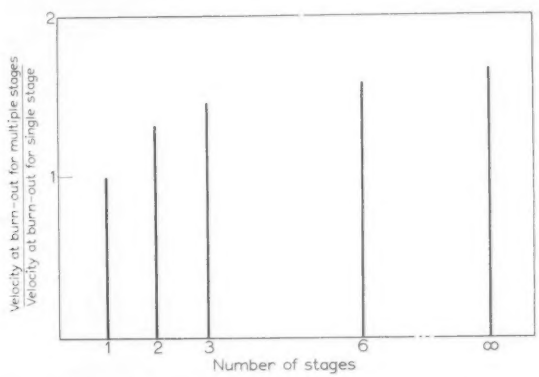


FIG. 3. Illustrating the effect of using a multistage rocket sequence. The velocity at burn-out, for a given fuel consumption, is shown as a function of the number of stages.

maximum speed attained, for a given fuel consumption, with the number of stages employed. It will be seen that the improvement gained as the number increases beyond 4 is relatively small and is likely to be offset by the increased complexity. The American vehicles which have been launched have either employed 4 (the Explorer type) or 3 stages (the Vanguard type).

It is instructive to note the considerations involved in the choice of the launching procedure and the guidance system for the Vanguard vehicles (HAGEN, 1957). An important further choice had to be made before deciding the performance requirement for the different stages. This concerned the stage in which the guiding mechanism is to be housed. The final stage is ruled out because of the need to carry the mass of this equipment to the maximum height. This led to the choice of the second stage to house the guidance as it was considered impracticable to carry through the whole sequence beyond the jettisoning of the first stage without active guidance. A total mass requirement of 22,600 lb was then arrived at, in order to launch a spherical satellite of 20 lb.

By choosing optimum weight and trajectories for the different stages the data given in Table 1 were obtained.

The launching sequence is then as follows. The ascent begins vertically under the power of the first stage rocket and gradually tilts in the required direction. The first stage burn-out occurs about 142 sec after launching at which time an altitude of 35 miles and a speed of 6000 ft/sec have been reached. The first stage then separates and the second stage ignites; so the vehicle proceeds on a more and more inclined

path to an altitude of 135 miles at burn-out of this stage. During this stage the speed is increased to 14,000 ft/sec. The second stage is not jettisoned at burn-out but coasts forward a distance of about 700 miles while rising to the final altitude of 300 miles. While coasting the vehicle is brought to the correct orientation for the

Table 1
Vanguard weight and performance data

	Stage		
	1	2	3
Thrust (lb)	27,000	7,500	2,350
Burning time (sec)	146	120	30
Horizontal velocity increment (ft/sec)	4,023	8,389	13,405

third stage projection and a spin is imparted to this stage so the orientation is maintained after the second stage, with the guiding mechanism, has been jettisoned. Once the required altitude and orientation have been attained the third stage is separated and fired so as to impart the necessary additional horizontal velocity.

Given this programme the choice of propellents may be made. The first stage, which provides most of the power necessary to take the second and third stages up to the launching height, may be based on the existing Viking rocket motor which uses liquid oxygen and gasoline as propellents. For the second stage a gas-pressurized motor employing nitric acid and unsymmetrical dimethyl-hydrazine is employed, this possessing certain advantages in simplicity. Finally, for the third stage, a solid propellent is desirable. With a solid fuel the mass requirement is lower and there is likely to be less risk of disturbance of orientation due to asymmetry in burning or to "sloshing" of liquids.

Guidance in the Vanguard system is fully automatic. It includes altitude controls for the first-stage flight and for the second-stage powered and coasting flight, a programmer to vary the pitch angle in accordance with optimum requirements and a coasting-time computer which, by integrating the acceleration, establishes the optimum time for ignition of the third rocket.

Much less detail is available about the launching sequence for the American Explorer satellites (see, however, EXPLORER I, 1958). In principle it does not differ very greatly from that for Vanguard except that the sequence control is not fully automatic, the ignition of the second stage being controlled from the ground to which telemetred information is transmitted during the flight. The first-stage motor is based on the Jupiter rocket using a hydrazine-based liquid fuel in place of alcohol. The overall weight of the vehicle is about 65,000 lb and the first stage rocket motor has a thrust of about 83,000 lb. A further difference from the Vanguard sequence is that the first-stage motor provides all the power necessary to take the vehicle up to the final altitude. Ignition of the second stage does not take place until the orientation has been guided accurately into the horizontal so that the function of this and later stages is to boost the horizontal velocity to the desired value.

Some discussion of the problem of optimizing trajectories has been published by SIRY (1957) and by OKHOTSIMSKII and ENEEV (1957). As an example of this kind of analysis, we follow the argument given by the latter authors for deducing the optimum rate of variation of pitch angle with time, for flight of a rocket under conditions in which aerodynamic forces may be neglected.

If the curvature of the Earth can also be neglected the equations of motion of the rocket can be written in the form

$$\frac{\mathrm{d}u}{\mathrm{d}t} = p\cos\phi, \quad \frac{\mathrm{d}v}{\mathrm{d}t} = p\sin\phi - g,$$

where u and v are the horizontal and vertical components of velocity, respectively. p is the acceleration due to the rocket thrust which is in a direction inclined at an angle ϕ to the horizontal. It may be assumed that the thrust is along the rocket axis so that ϕ is also the inclination of this axis to the horizontal.

Both p and ϕ are to be regarded as functions of the time which may be chosen to achieve certain optimum conditions. When a chosen height h is reached, the vertical component of velocity is to be zero and the mass bears a given ratio to the initial mass. The horizontal component of velocity at the height h is to be a maximum consistent with these conditions. The time T to reach the altitude h may either be given a chosen value or be optimized.

This is an isoperimetric variational problem which may be solved as follows. It is convenient first to introduce a velocity V such that

$$V = \int_0^t p\,\mathrm{d}t, \; p = \frac{\mathrm{d}V}{\mathrm{d}t}.$$

The horizontal velocity after time T is

$$u(T) = \int_0^T \frac{\mathrm{d}V}{\mathrm{d}t}\cos\phi\,\mathrm{d}t + u_0,$$

where u_0 is the initial horizontal velocity. This is to be maximized consistent with the equations

$$\frac{\mathrm{d}v}{\mathrm{d}t} - \frac{\mathrm{d}V}{\mathrm{d}t}\sin\phi + g = 0,$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} - v = 0,$$

and the boundary conditions

$$u = u_0, \; v = v_0, \; y = y_0, \; V = 0; \; t = 0;$$

$$v = 0, \; y = h, \; V = V_f, \; t = T.$$

The magnitude of V_f is determined by the given ratio of the initial and final masses of the rocket.

Following the usual method of undetermined multipliers we require that

$$I = u(T) + \int_0^T \left\{\lambda_1\left(\frac{\mathrm{d}v}{\mathrm{d}t} - \frac{\mathrm{d}V}{\mathrm{d}t}\sin\phi + g\right) + \lambda_2\left(\frac{\mathrm{d}y}{\mathrm{d}t} - v\right)\right\}\mathrm{d}t$$

have an extreme value with respect to variations δy, δv, δV, $\delta\phi$ and δT, which may be treated as independent.

Since

$$\delta I = g\lambda_1\delta T + \int_0^T \Big[\Big(-\sin\phi - \lambda_1\cos\phi\Big)\frac{\mathrm{d}V}{\mathrm{d}t}\delta\phi + \\ + \Big\{-\frac{\mathrm{d}}{\mathrm{d}t}\Big(\cos\phi - \lambda_1\sin\phi\Big)\Big\}\delta V - \frac{\mathrm{d}\lambda_2}{\mathrm{d}t}\delta y - \Big(\frac{\mathrm{d}\lambda_1}{\mathrm{d}t} + \lambda_2\Big)\delta v\Big]\mathrm{d}t,$$

we must have

$$\sin\phi + \lambda_1\cos\phi = 0, \qquad \ldots\ldots (2.12)$$

$$\frac{\mathrm{d}\lambda_1}{\mathrm{d}t} = -\lambda_2, \quad \frac{\mathrm{d}\lambda_2}{\mathrm{d}t} = 0. \qquad \ldots\ldots (2.13)$$

Hence

$$\lambda_1 = C_1 + C_2 t,$$

where C_1 and C_2 are constants, so that from (2.12)

$$\tan\phi = -(C_1 + C_2 t). \qquad \ldots\ldots (2.14)$$

If T is not fixed we must also have $\lambda_1(T) = 0$ so that

$$\tan\phi = C_2(T - t). \qquad \ldots\ldots (2.15)$$

The equations (2.14), (2.15) show that the pitch angle ϕ should be controlled so its tangent varies linearly with the time. This result is independent of the form of the function, V, so it applies if the launching rocket is a multistage one. The function V may be given or it may be optimized.

Okhotsimskii and Eneev (1957) have applied these considerations to multistage launching rockets for which the motion takes place at constant thrust and have given numerical details. They also consider the modifications which have to be introduced into the argument if the curvature of the Earth is allowed for. Provided the horizontal distance traversed from the launching site to the attainment of the desired altitude and horizontal velocity is small compared with the radius of Earth, the linear control formulae (2.14) and (2.15) remain good approximations.

3. Perturbations of Satellite Orbits

There are two main sources of perturbation of the orbits of artificial satellites—the air resistance and the oblateness of the Earth. We shall consider these in turn.

Air Resistance

There are two sources of drag, one due to momentum interchange with neutral atmospheric atoms and molecules, the other due to the effect of the ionosphere in charging up the satellite so that there is an additional momentum interchange with the atmospheric ions.

Drag due to Neutral Molecules

The drag force on a body moving with high speed V through a gas is usually written in the form $\frac{1}{2}C\rho V^2S$ where S is the effective cross-section of the body, ρ the density of the gas and C the drag coefficient. We consider a spherical satellite of radius R. It is possible to relate C to the transfer of momentum between the gas atoms or molecules and the satellite in terms of simple gas-kinetic considerations.

At the altitude of a satellite orbit the free path of the gas molecules is large compared with the satellite dimensions, and in addition the speed V of the satellite is large compared with the mean speed $\bar{v}$ of the molecules. Thus for air at temperature 1000°K, $\bar{v} = 10^5$ cm/sec whereas $V = 8 \times 10^5$ cm/sec. If the impacts between gas molecules and a satellite are elastic the rate of momentum transfer to the gas is given by

$$F = \pi R^2 nmV^2, \qquad \ldots . (3.1)$$

where n is the number of gas molecules/cm^3 and m is the mass of a molecule. This assumes that the molecules are specularly reflected on impact. If, however, they spend some time on the surface before re-emission, we have the case of diffuse reflection and

$$F = \frac{4}{3}\pi R^2 nmV^2. \qquad \ldots . (3.2)$$

We then have $C = 2$ or $8/3$ in cases (3.1), (3.2) respectively. It is difficult to see how a larger value of C can arise for a spherical satellite due to collisions with neutral molecules. If the collisions are partly inelastic C will be smaller.

Electrical Drag

The possibility of a significant contribution to the drag due to charging up of the satellite has been discussed by JASTROW and PEARSE (1957). They consider in the first instance the potential which the satellite will acquire relative to its surroundings due to the different mobilities of the positive ions and electrons of the ionosphere through which it passes.

The potential will adjust itself so that the net inward flux of electrons equals that of the positive ions. In daytime, allowance must be made for photoelectric emission from the satellite but at night the problem is simpler. Because of the higher mobility of the electrons the equilibrium potential ϕ_0 will then be negative to space. Hence if the velocity distribution of the electrons is Maxwellian with a temperature T_e the total number of electrons reaching the satellite per second will be

$$\pi R^2 n_e (2mk^2T_e^2)^{-\frac{1}{2}} \int_{e\phi_0}^{\infty} E^{\frac{1}{2}} e^{-E/kT_e} dE, \qquad \ldots . (3.3)$$

where n_e is the number of electrons/cm^3.

The number of positive ions reaching the satellite will be given by

$$\pi R^2 n_i V, \qquad \ldots . (3.4)$$

where V is the satellite velocity and n_i the number of ions/cm^3; ϕ_0 is obtained by equating (3.3) and (3.4) and putting $n_e = n_i$.

The satellite potential then comes out to be a few times kT_e/e. It is difficult to decide what is the correct value to take for T_e but when the satellite is in darkness it is probably not much above the temperature of the neutral molecules. In that case ϕ_0 would not be more than a few volts (see p. 102). JASTROW and PEARSE take a somewhat higher value on the assumption that T_e is about 15 times the gas temperature.

It must be remembered also that ϕ_0 is very sensitive to the shape of the electron energy distribution. Thus if there is a "tail" to the distribution, of comparatively high energy electrons, the contribution to the incoming current due to these electrons will be relatively so high that the satellite must charge up to a high potential to prevent the faster electrons from reaching the surface. In daytime there may well exist a relatively large proportion of electrons with energies as high as 50 eV or more. These will be the electrons produced by photoionization of atoms and molecules by the short wavelength radiation from the solar corona. It is not possible to say from this that a satellite will charge up to potentials as high as 50 V in daylight because photoelectric emission, rather difficult to estimate, will act to reduce this potential.

JASTROW and PEARSE have calculated the increased target area of a charged satellite towards collisions with positive ions. Due to the effect of the negative charge the orbits of the positive ions are modified so that the largest impact parameter which an ion may have in order to hit the satellite's surface is given by

$$p = R + \alpha\lambda,$$

where $\lambda = (kT_e/4\pi n_e e^2)^{\frac{1}{2}}$ is the Debye length which determines the scale of the polarization shielding round the satellite, and α increases from 6 to 10 as the satellite potential increases from 10 to 60 V. For $kT_e = 1{\cdot}5$ eV, $n_e = 2\times10^5/\text{cm}^3$, $\lambda = 2$ cm so that $\alpha\lambda$ is comparable with R for a satellite of 20 in. diameter charged to a few tens of volts. This increases the effective cross section of the satellite for momentum transfers with positive ions.

Figure 4 illustrates the ratio of charged to neutral drag calculated by JASTROW and PEARSE for a satellite at an altitude of about 600 km charged to different negative

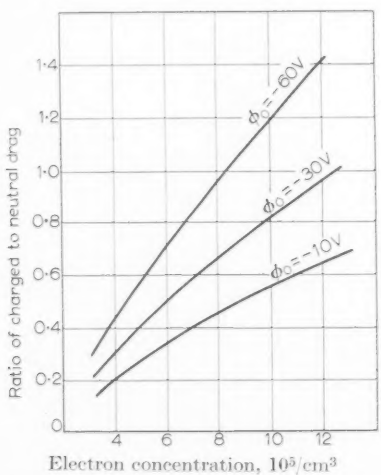


FIG. 4. Ratio of the charged to the neutral drag for a spherical satellite, charged to different negative potentials ϕ_0, as a function of electron concentration. The neutral molecule concentration is taken to be $10^7/\text{cm}^3$ and the electron temperature T_e as given by $kT_e = 1{\cdot}5$ eV.

potentials, kT_e being taken as 1·5 eV, the concentration of neutral particles being assumed to be $10^7/\text{cm}^3$ and of the electrons up to $1{\cdot}2 \times 10^6/\text{cm}^3$.

Under these conditions the contribution from the charge drag is appreciable but it is by no means certain that the magnitudes of the potential acquired will be as high as 30 V. It is likely that in Fig. 4 the contribution from the charge drag is overestimated. At lower altitudes it will be even less important. Thus at 200 km the neutral concentration will be about 1000 times larger whereas the maximum electron concentration at any altitude will not be larger than the maximum assumed in the preparation of Fig. 4.

It seems that, for altitudes below 600 km, it is safe to ignore charge drag and an upper limit to the drag coefficient can be taken as 8/3.

The Effect of Air Drag on the Orbit

The effect of a small tangential disturbing force on an elliptic orbit has been discussed fully in astronomical texts (MOULTON 1931, SMART 1953) and is also a familiar part of the orbit theory which appears in texts on particle dynamics.

A tangential impulse T applied when the body is at a distance r from the centre will produce an impulsive change in the magnitude of the velocity V but no impulsive change in r. From the relation (2.1) we have that the change δa of the semi-major axis is given by

$$\delta a = \frac{2Va^2}{\mu}\delta V. \qquad \ldots . (3.5)$$

The air drag which acts to reduce V will thus produce a reduction in the major axis, and hence in the period τ of revolution which is given by

$$\delta\tau/\tau = \frac{3}{2}\,\delta a/a. \qquad \ldots . (3.6)$$

Referring to (2.2) we see that the change δe in the eccentricity is given by

$$\delta a(1-e^2) - 2ea\delta e = 2V\delta V r^2 \cos^2\theta/\mu,$$

so, making use of (3.5)

$$\delta e = -\frac{V\delta V}{e\mu a}\{r^2\cos^2\theta - a^2(1-e^2)\}, \qquad \ldots . (3.7)$$

$$= -\frac{V\delta V}{e\mu a}\{r^2\cos^2\theta - b^2\}, \qquad \ldots . (3.8)$$

where $2b$ is the minor axis of the instantaneous ellipse. It will be seen from (3.8) that δe vanishes when the impulse is applied at the end of the minor axis. For a retarding impulse ($\delta V < 0$) δe is negative for an impulse applied on the side of the minor axis containing the occupied focus, positive on the other side. As the retarding force is proportional to v^2 it is clear that the integrated effect on the latter side will be less than that on the former so that the air drag will produce a reduction of the eccentricity each revolution.

The line of apsides will not undergo a secular rotation due to air drag, but there will be an oscillating effect. Because the direction of the velocity is not changed by a tangential impulse the direction of the line to the unoccupied focus will be unchanged. However, if r' is the distance from this focus $r' = 2a - r$ so that r' is reduced. Referring to Fig. 5 we see then that a tangential impulse will produce equal and opposite

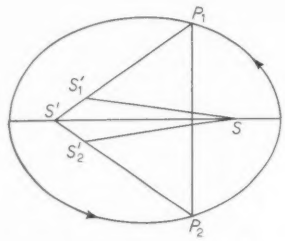


FIG. 5. Illustrating the effect of tangential perturbing force on the apse line of an elliptic orbit. A retarding impulse at P_1 displaces the unoccupied focus S' to S_1', an equal impulse at P_2 displaces it to S_2'.

rotation of the apse line when applied at points symmetrically placed with respect to the major axis. Hence the air drag will produce a negative rotation of the apse line when the upper half of the orbit is being followed and a nearly equal and opposite rotation in the lower half.

Summarizing, we can say that the effect of air drag will be to reduce the major axis and decrease the eccentricity so that the orbit becomes more and more nearly circular. The period will decrease steadily but there will be no appreciable mean rotation of the line of apsides.

There is no difficulty in carrying the analysis further so that quantitative information about the air drag can be derived from observations of the change of period as a function of the number of revolutions.

In terms of polar co-ordinates r and ϑ in the plane of motion with occupied focus as pole and the polar axis along the major axis directed away from the occupied focus (Fig. 6), we have

$$r = \frac{a(1-e^2)}{1+e\cos\vartheta}, \quad r^2\dot{\vartheta} = \{\mu a(1-e^2)\}^{\frac{1}{2}}, \qquad \ldots . (3.9)$$

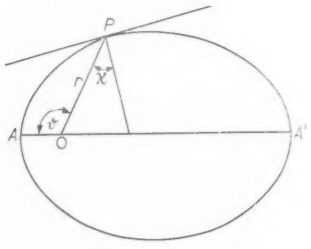


FIG. 6.

so that

$$(1-e^2)^{\frac{3}{2}}(1+e\cos\vartheta)^{-2}\dot{\vartheta} = (\mu/a^3)^{\frac{1}{2}} \qquad \ldots . (3.10)$$

and from (2.1)

$$aV^2 = \mu(1+2e\cos\vartheta+e^2)(1-e^2)^{-1}. \qquad \ldots\ldots (3.11)$$

Since from (3.5) we have

$$\begin{aligned}\frac{da}{d\vartheta} = \frac{1}{\dot\vartheta}\frac{da}{dt} &= \frac{2}{\dot\vartheta}\frac{Va^2}{\mu}\frac{dV}{dt},\\ &= \frac{2}{\dot\vartheta}\frac{Va^2}{\mu}T,\end{aligned} \qquad \ldots\ldots (3.12)$$

where T is the tangential perturbing force per unit mass, it follows from (3.12) that

$$\frac{da}{d\vartheta} = \frac{2a^3(1-e^2)}{\mu(1+e\cos\vartheta)}(1+2e\cos\vartheta+e^2)^{\frac{1}{2}}T. \qquad \ldots\ldots (3.13)$$

With the law of resistance discussed in §3, p. 39,

$$T = -\tfrac{1}{2}\kappa\rho V^2, \qquad \ldots\ldots (3.14)$$

where $\kappa = CS/m$, m being the mass of the satellite.
Hence, from (3.13)

$$\frac{da}{d\vartheta} = -\frac{a^2}{(1+e\cos\vartheta)^2}(1+2e\cos\vartheta+e^2)^{\frac{1}{2}}\kappa\rho. \qquad \ldots\ldots (3.15)$$

We also note for further reference that, by using (3.8) and (3.14),

$$\frac{de}{d\vartheta} = -\kappa\rho a(1-e^2)(e+\cos\vartheta)(1+2e\cos\vartheta+e^2)^{\frac{1}{2}}/(1+e\cos\vartheta)^2, \qquad \ldots\ldots (3.16)$$

so

$$\frac{d}{d\vartheta}\{a(1\pm e)\} = -\kappa\rho a^2(1\pm e)^2(1\pm\cos\vartheta)(1+2e\cos\vartheta+e^2)^{\frac{1}{2}}/(1+e\cos\vartheta)^2 \qquad \ldots\ldots (3.17)$$

The change in the major axis in a period of revolution is then

$$\Delta a = -a^2\kappa\int_0^{2\pi}(1+2e\cos\vartheta+e^2)^{\frac{1}{2}}(1+e\cos\vartheta)^{-2}\rho d\vartheta, \qquad \ldots\ldots (3.18)$$

and from (3.6) the required rate of change of period per revolution is given by

$$\frac{\Delta\tau}{\tau} = -\frac{3}{2}a\kappa\int_0^{2n}(1+2e\cos\vartheta+e^2)^{\frac{1}{2}}(1+e\cos\vartheta)^{-2}\rho d\vartheta. \qquad \ldots\ldots (3.19)$$

It is of course assumed in all this analysis that $\Delta\tau/\tau$ is very small compared with unity.

The density ρ falls off rapidly with altitude so that the main contribution to the change of a or τ comes from those parts of the orbit at and near the perigee. We may represent the variation of ρ with height h in this region by the exponential form

$$\rho = \rho_0 \exp\{-(h-h_0)/H\} \quad \ldots . (3.20)$$

where ρ_0 is the density at the perigee, which is at an altitude h_0, and H is the scale-height. Provided

$$H \ll l - r_p, \quad \ldots . (3.21)$$

where l is the semi-latus rectum ($\vartheta = \pi/2$) and $r_p(\vartheta = 0)$ the distance of the perigee from the centre of the Earth, we can regard the contribution to the change of a in a revolution to be confined to such small values of ϑ that an expansion in powers of ϑ will give rapid convergence. The condition (3.21) may be written

$$H \ll ae(1-e). \quad \ldots . (3.22)$$

When this is satisfied we write, following GROVES (1959)

$$(1+2e\cos\vartheta+e^2)^{\frac{3}{2}}(1+e\cos\vartheta)^2 = (1+e)(1+\lambda_2\vartheta^2+\lambda_4\vartheta^4+\ldots) \quad \ldots . (3.23)$$

where

$$\lambda_2 = -e(1-2e)/2(1+e)^2, \quad \ldots . (3.24)$$

$$\lambda_4 = e(1-9e-3e^2+16e^3)/24(1+e)^4. \quad \ldots . (3.25)$$

Also, if R is the radius of the Earth,

$$h = \frac{a(1-e^2)}{1+e\cos\vartheta} - R,$$

$$h_0 = a(1-e),$$

so

$$\rho = \rho_0 \exp -\{\alpha_2\vartheta^2+\alpha_4\vartheta^4+\ldots\}, \quad \ldots . (3.26)$$

where

$$\alpha_2 = H^{-1}\ \frac{ae(1-e)}{2(1+e)}, \qquad \alpha_4 = -H^{-1}\ \frac{a\,e(1-e)(1-5e)}{24(1+e)^2}. \quad \ldots . (3.27)$$

Hence

$$\frac{da}{d\vartheta} = -a^2\kappa\rho_0(1+e)\exp\{-(\alpha_2\vartheta^2+\alpha_4\vartheta^4+\ldots)\}(1+\lambda_2\vartheta^2+\lambda_4\vartheta^4+\ldots) \quad \ldots . (3.28)$$

To evaluate the change of a over a complete revolution we may integrate over a range of ϑ from $-\vartheta_0$ to $+\vartheta_0$ where ϑ_0 is such that the contribution from larger values of ϑ is negligible. In practice, provided (3.22) is well satisfied the limits of

integration can be taken as $\pm\infty$ and the change Δa in a revolution is given to a close approximation by

$$\Delta a = -a^2\kappa\rho_0(1+e)\int_{-\infty}^{\infty} e^{-\alpha_2\vartheta^2}[1+\lambda_2\vartheta^2+(-\lambda_4+\alpha_4)\vartheta^4]\mathrm{d}\vartheta,$$

$$= -a^2\kappa\rho_0(1+e)(\pi/\alpha_2)^{\frac{1}{2}}(1+\epsilon/\alpha_2), \qquad \ldots\ldots (3.29)$$

where

$$\epsilon = -\frac{1}{2}\left(-\lambda_2+\frac{3}{2}\frac{\alpha_4}{\alpha_2}-\frac{3}{2}\frac{\lambda_4}{\alpha_2}\right).$$

The change $\Delta\tau$ in period per revolution may be obtained from the relation

$$\frac{\Delta\tau}{\tau} = \frac{3}{2}\frac{\Delta a}{a}, \qquad \ldots\ldots (3.30)$$

so that

$$\frac{\Delta\tau}{\tau} = -\frac{3}{2}a\kappa\rho_0(1+e)(\pi/\alpha_2)^{\frac{1}{2}}(1+\epsilon/\alpha_2). \qquad \ldots\ldots (3.31)$$

This is a convenient formula to use in practice, at least in the earlier phase of a satellite's life when the orbit is sufficiently eccentric for the condition (3.22) to be well satisfied. It has been generalized to take account of the oblateness of the atmosphere and the fact that the scale height varies with altitude (GROVES 1959). Information about air density obtained by application of the formula to the motion of the first artificial satellites is discussed in §13, p. 97.

In the later phase of a satellite's life the orbit becomes nearly circular and it is necessary to employ a different procedure (FEJER 1957, SCOTT 1957, LESLIE 1958). Thus if the eccentricity is so small that

$$ae^2/H \ll 1 \qquad \ldots\ldots (3.32)$$

then, following YATSUNSKII (1957), we write

$$h-h_0 \simeq ae(1-\cos\vartheta)-ae^2(1-\cos^2\vartheta),$$

$$(1+2e\cos\vartheta+e^2)^{\frac{3}{2}}(1+e\cos\vartheta)^{-2} \simeq 1+e\cos\vartheta,$$

so that

$$\frac{\Delta\tau}{\tau} \simeq -\frac{3}{2}a\kappa\rho_0 e^{-ae/H}\int_0^{2\pi}\exp\{ae\cos\vartheta/H\}\left(1+e\cos\vartheta+\frac{ae^2}{H}(1-\cos^2\vartheta)\right)\mathrm{d}\vartheta$$

$$\simeq -3\pi a\kappa\rho_0 e^{-ae/H}\left[I_0\left(\frac{ae}{H}\right)+eI_1\left(\frac{ae}{H}\right)\right]. \qquad \ldots\ldots (3.33)$$

where I_0 and I_1, are the usual Bessel functions.

A useful formula for the lifetime of a satellite after the eccentricity of its orbit has fallen to a value which satisfies (3.32) may be obtained as follows. We measure the

number of revolutions from an initial time at which (3.32) is valid, and only retain terms of lowest order in e. Writing r_a, r_p for the apogee and perigee distances at any subsequent time we then have from (3.17) for the changes per revolution,

$$\Delta r_a = -2\pi\kappa e_0 a^2 e^{-ae/H}\left[I_0\left(\frac{ae}{H}\right)+I_1\left(\frac{ae}{H}\right)\right] \quad \ldots . (3.34)$$

$$\Delta r_p = -2\pi\kappa e_0 a^2 e^{-ae/H}\left[I_0\left(\frac{ae}{H}\right)-I_0\left(\frac{ae}{H}\right)\right] \quad \ldots . (3.35)$$

in which the suffix $_0$ distinguishes the initial values of the different quantities. Since $r_a+r_p = 2a$, $r_a-r_p = 2ac$ we have, again including only terms of lowest order in e,

$$\Delta a = -2\pi a^2\kappa e_0 e^{-ae/H} I_0\left(\frac{ae}{H}\right) \quad \ldots . (3.36)$$

$$\Delta(ae) = -2\pi a^2\kappa e_0 e^{-ae/H} I_1\left(\frac{ae}{H}\right) \quad \ldots . (3.37)$$

Hence

$$\Delta(ae) = \frac{I_1(ae/H)}{I_0(ae/H)}\Delta a.$$

Since

$$zI_0(z) = \frac{\mathrm{d}}{\mathrm{d}z}[zI_1(z)],$$

we have

$$\Delta a = H\Delta[\log\{aeI_1(ae/H)\}]$$

so that, integrating,

$$\exp[H^{-1}\{a_0-a\}] = \frac{a_0e_0I_1(a_0e_0/H)}{aeI_1(ae/H)}. \quad \ldots . (3.38)$$

On substitution in (3.37) and remembering that, to first order in e, $a^2 = a_0{}^2$, we have

$$(e_0{}^2-e^2) = 4\pi\kappa a_0\rho_0 e_0 \exp(-a_0e_0/H)I_1(a_0e_0/H)N, \quad \ldots . (3.39)$$

where N is the number of orbital revolutions since the initial time.

Provided the initial perigee height is not too great the number of revolutions made by the satellite after the orbit becomes circular will be small. It follows that the number of revolutions remaining after the initial time is given by

$$N = e_0/[4\pi\kappa a_0e_0e^{-a_0e_0/H}I_1(a_0e_0/H)].$$

If the initial conditions are such that $a_0e_0/H \gg 1$ while $a_0e_0{}^2/H \ll 1$ then $I_1(a_0e_0/H) \simeq$ $\simeq I_0(a_0e_0/H)$ and we see from (3.36) and (3.30) that

$$N = -\tfrac{3}{4}e_0\tau_0/(\Delta\tau)_0. \quad \ldots . (3.40)$$

The lifetime t_L of the satellite is then given to order e_0 by

$$\begin{aligned} t_L &= N\tau_0 \\ &= -\tfrac{3}{4}e_0\tau_0/(\Delta\tau/\tau)_0 \end{aligned} \quad \ldots . (3.41)$$

KING-HELE and LESLIE (1958) have extended this analysis to apply when the condition (3.32) is no longer satisfied. They find that (3.42) becomes

$$t_L = -\tfrac{3}{4}\{e_0\tau_0/(\Delta\tau/\tau)_0\}\left\{1+\frac{7}{6}e_0+\frac{H}{2a_0e_0}+O(e_0^2)+O\left(\frac{H^2}{a_0^2e_0^2}\right)\right\}. \quad \ldots . (3.42)$$

This shows that, in practice, (3.41) remains a good approximation even though (3.32) is not satisfied.

A further useful formula, also obtained by KING-HELE and LESLIE (1958) gives the orbital period, at time t after launching, in the form

$$\frac{\tau}{\tau_0} = 1-\tfrac{3}{2}e_0(1-\tfrac{5}{3}e_0)[1-\sqrt{\{1-(t/t_L)\}}]-\tfrac{17}{8}e_0^2\frac{t}{t_L}+\frac{3H}{8a_0}\log[1-(t/t_L)]+$$

$$+O(e_0^3)+O(H^2/a_0^2),$$

which, for practical cases, reduces without serious error to

$$\tau/\tau_0 = 1-\tfrac{1}{2}(3e_0)[1-\sqrt{\{1-(t/t_L)\}}]. \quad \ldots . (3.43)$$

The effect of the oblateness of the atmosphere in this type of analysis has been considered by PARKYN (1958a and b).

A detailed pre-launching study of the effect of air drag was made by OKHOTSIMSKY, ENEEV and TARATYNOVA. They represented the height variation of density in the form

$$\rho = \rho_0\chi\left(1+\frac{h-h_0}{\alpha}\right)^{-k} \quad \ldots . (3.44)$$

where the values given in Table 2 were taken for the constants χ, α, h_0 and k in different height ranges.

Table 2

Altitude h	χ	α	h_0	k
km		km	km	
100–150	1	55	100	8
150–250	$5{\cdot}667\times10^{-3}$	100	150	7
250–900	$4{\cdot}428\times10^{-3}$	215	250	6

These were chosen to give close agreement with the atmospheric model given in MITRA's book on the upper atmosphere. ρ_0, the density at 100 km, was taken as $1{\cdot}73\times$ $\times10^{-9}$ g/cm^3.

The procedure consisted in determining the number of revolutions made by a satellite launched in an orbit with given apogee and perigee, until the perigee distance is reduced to 100 km after which its lifetime was assumed to be negligible. This involved numerical solutions of the coupled equations for the rates of change of semi-latus rectum l and eccentricity e with number of revolutions.

Although the detailed results now need modification in the light of data obtained from the first few satellites they are still of interest as providing useful indication of what is to be expected. Figure 7 is a reproduction of these results. A curve such as

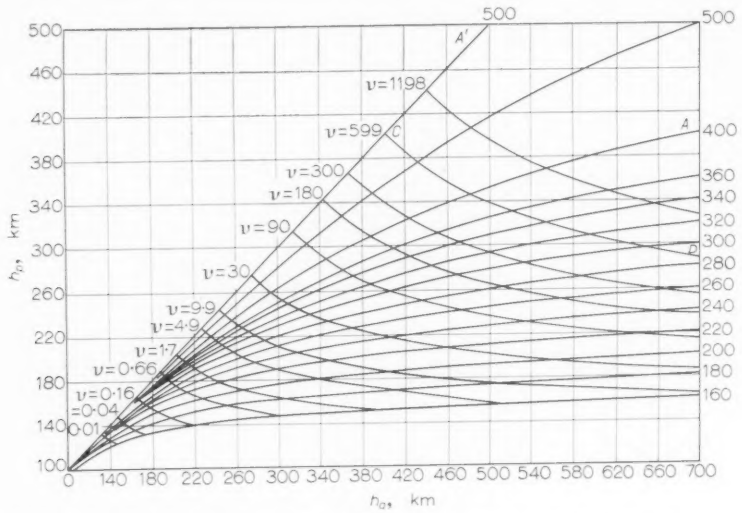


FIG. 7. Diagram for estimating the number of N of revolutions made by a satellite with given initial apogee (h_a) and perigee (h_p) heights, before the perigee height falls to 100 km N is given in terms of ν by

$$N = m\nu/9{\cdot}81\,CS$$

where m is the satellite's mass in kg, S its effective cross sectional area in m^2 and C the drag coefficient.

OA represents a relation between apogee and perigee height which enables one to trace the variation in these quantities from any given pair of initial values. Thus OA' corresponds to circular orbits. The family of curves such as OA is intersected by a second set of which CD is a typical member. These are curves on which a certain parameter ν is constant. ν is given by

$$\nu = 9{\cdot}81NCS/m$$

where C is the drag coefficient, S the cross section in m^2 and m the mass in kg of the satellite. N is the number of revolutions before the perigee reaches the 100 km altitude. Thus the values of ν at two points on a curve of the first family gives the number of revolutions made between the two orbits characterized by the particular apogee and perigee heights. Thus if we take a satellite of diameter 20 in. and mass 84 kg, and $C = 2$ we have

$$N = 22{\cdot}8\,\nu.$$

Thus the number of revolutions made by the satellite, launched with apogee and perigee distances 660 and 220 km respectively, before the perigee height fell to 100 km would be $90 \times 22{\cdot}8 = 2052$ (128 days).

The rapid decrease of apogee distance compared with that of perigee is very clear in Fig 7.

Pre-launching calculations of the effects of air resistance were also carried out by other authors. Results of a calculation due to TARATYNOVA (1957) in which the

effects of the Earth's oblateness were included as well as that of air resistance, for a particular case, are illustrated in Fig. 11. DAVIS, WHIPPLE and ZIRKER (1957) have considered, in addition, the effect of air drag on the spin of a satellite and find it to be very small.

The inverse problem of determining the atmospheric density from the observed rate of change of period of revolution with time is the most important from the geophysical viewpoint. From a formula such as (3.31), where it is applicable, the density at the perigee height may be obtained from the observed value of $\Delta\tau/\tau$ provided the scale height H is assumed. As the perigee distance changes it is possible in principle to determine the scale height also. In practice, while the orbit is sufficiently elliptical for (3.23) to be valid, the change of perigee distance is so small that it is difficult to obtain the necessary accuracy in the measurement of the perigee height. GROVES (1958 and 1959) has pointed out, however, that under certain circumstances advantage may be taken of the fact that the perigee latitude changes during the motion due to the spheroidal shape of the Earth (see §3, p. 55). As a result of this the perigee altitude, reckoned from the surface, may change by amounts comparable with the difference between the polar and equatorial radii, 13 miles. This in turn is comparable with the scale height so that the rate of change of period may alter very significantly.

The perigee height when the satellite is at a geocentric latitude ψ is given by

$$h = a(1-e) - R(\psi),$$

where $R(\psi)$ is the Earth's radius at latitude ψ. Assuming the spheroidal shape we have

$$R(\psi) = R_0 - \Delta R \sin^2\psi,$$

where R_0 and $R_0 - \Delta R$ are the equatorial and polar radii respectively.

Hence by (3.20), if suffixes 1 and 2 denote values at different epochs,

$$\frac{\rho_{01}}{\rho_{02}} = \exp[\{a_1(1-e_1) - a_2(1-e_2)\} - \Delta R(\sin^2\psi_1 - \sin^2\psi_2)\}/H] \quad \ldots . (3.45)$$

By applying the analysis of p. 44 to (3.17) we find that the change of $a(1-e)$ in a revolution is given by

$$\Delta\{a(1-e)\} = -\tfrac{1}{4}\kappa a^2\rho_0(1-e^2)(1+e)^{-1}(\pi/\varkappa_2{}^3)^{\frac{1}{2}}(1+\epsilon'/\varkappa_2)$$

where

$$\epsilon' = -\frac{3}{2}\left(\varkappa_2{}' + \frac{5}{2}\varkappa_4{}'/\varkappa_2\right),$$

with

$$\varkappa_2{}' = -\left(\mu_2 - \frac{1}{12}\right),$$

$$\varkappa_4{}' = \varkappa_4 - \mu_4 + \frac{1}{12}\mu_2 - \frac{1}{360}.$$

It follows from (3.29) that, to order α_2^{-2},

$$\frac{\mathrm{d}\{a(1-e)\}}{\mathrm{d}a} = \tfrac{1}{4}(1-e)^2(1+e)^{-2}[1+(\epsilon'-\epsilon)\alpha_2]\alpha_2^{-1}.$$

As a first approximation, ignoring terms of order α_2^{-1}

$$a(1-e) = a_0(1-e_0). \qquad \ldots . (3.46)$$

To the same accuracy

$$\frac{\mathrm{d}}{\mathrm{d}a}\left\{\frac{ae(1-e)}{1+e}\right\} = \frac{(1-e)^2}{(1+e)^2},$$

so, by (3.27),

$$(1-e)^2(1+e)^{-2} = 2H\mathrm{d}\alpha_2/\mathrm{d}a,$$

and as a next approximation to (3.46)

$$a(1-e) = a_0(1-e_0) - \tfrac{3}{2}H\log(\alpha_{20}/\alpha_2) \qquad \ldots . (3.47)$$

It follows by substitution of (3.47) in (3.45) that the scale height H can be determined from the relation

$$H = [R(\psi_1)-R(\psi_2)]/\log(\rho_{01}\alpha_{22}^{\frac{1}{2}}/\rho_{02}\alpha_{21}^{\frac{1}{2}}), \qquad \ldots . (3.48)$$

ρ_{01} and ρ_{02} being determined from (3.31) using observed data.

In deriving (3.48) we have ignored the effects of the spheroidal shape of the Earth and the variation of the scale height with altitude. These have been included by Groves (1959) but will usually be small effects.

The application of the considerations of this section to the analysis of the effect of air drag on the satellites which have been launched will be discussed in §13, p. 99.

The Effect of the Polar Flattening of the Earth

So far, we have only considered the effect of a perturbing force in the plane of the orbit and directed along the tangent at each point. The perturbations due to the polar flattening have components perpendicular to the plane of the orbit and in both tangential and normal directions in the plane. It is convenient to adopt more definitely the standard approach in terms of orbital elements although we shall derive the necessary formulae by elementary methods.

As before we consider at any point the instantaneous ellipse which would be followed if all perturbations ceased at points beyond. We choose a right-handed set of reference axes with origin at the centre of the Earth, z-axis towards the North pole and x-axis along the equator. It is convenient, when discussing the orientation of the orbit, to work in terms of the projections of the instantaneous ellipse on the celestial sphere centred round the centre of the Earth.

The inclination i of the orbit is the angle the plane of the ellipse makes with the equatorial plane. The points of intersection of the orbit with the equatorial plane are the nodes, the ascending node being the one from which the motion is upward (in the sense of positive z). The angle the line of nodes makes with Ox, the longitude

of the ascending node, is usually denoted by Ω. In addition there is a third angular element ω which is the angular distance between the perigee and the ascending node (measured between the projections in the celestial sphere).

Referring to Fig. 8, NPQ is the projection on the celestial sphere of an arc of the

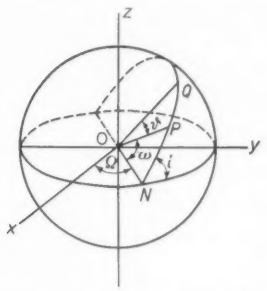


FIG. 8. Orbital angular elements. P is the projection of the perigee on the unit sphere.

ellipse. M is the ascending node, P the perigee point and Q a running point on the ellipse. The usual polar angle defining the position of Q is the angle PCQ or ϑ. It is convenient also to introduce the angle $u = \omega + \vartheta$ which is the polar angle of Q in the plane of the ellipse measured from the ascending node.

A change in Ω corresponds to rotation of the orbital plane round the z-axis while a change in ω corresponds to rotation of the orbit in its own plane.

Changes in Angular Elements due to Perturbations

Orthogonal perturbation.—A perturbing impulse perpendicular to the plane of the orbit will change all three angular elements. The primary effect is a change in direction of the velocity vector at the point where the impulse is applied. Considering again the projections on the celestial sphere we have the situation shown in Fig. 9.

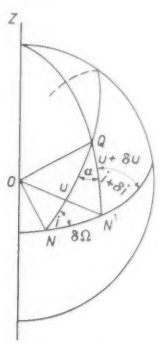


FIG. 9.

At Q the impulse produces a deflection of the velocity vector through an angle α as projected on the celestial sphere. This changes the plane of the orbit from NOQ to $N'OQ$. Thus in the spherical triangle NQN' (Fig. 9) the angles and sides are as indicated. It follows that

$$\frac{\sin \delta\Omega}{\sin \alpha} = \frac{\sin(u+\delta u)}{\sin i}$$

so, as $\delta\Omega$, δu and α are small,

$$\delta\Omega = \alpha\frac{\sin u}{\sin i}.$$

Also, by applying DELAMBRE's analogy,

$$\tan\tfrac{1}{2}(A+B) = \frac{\cos\frac{1}{2}(a-b)}{\cos\frac{1}{2}(a+b)}\cos\tfrac{1}{2}C,$$

with $A = i$, $B = \pi-(i+\delta i)$, $C = \alpha$, $a = u+\delta u$, $b = u$ we find, to the first order,

$$\delta i = \alpha\cos u.$$

Finally, to obtain $\delta\omega$ we use the fact that the z-coordinate of Q is

$$r\sin(\vartheta+\omega)\sin i,$$

where r is the distance OQ. Since $\delta z = 0 = \delta r$ we must have

$$\delta\omega\cos(\vartheta+\omega)\sin i+\delta i\sin(\vartheta+\omega)\cos i = 0,$$

so

$$\begin{aligned}\delta\omega &= -\tan u\cot i\delta i,\\ &= -\alpha\sin u\cot i.\end{aligned}$$

α is the angle between the initial and final directions of motion of the point Q as projected on the celestial sphere. Hence, if δV is the change of velocity due to the impulse

$$\alpha = \delta V/V\cos\chi,$$

where $V\cos\chi$ is the component of velocity normal to the radius vector. Since $rV\cos\chi$ is the angular momentum in the orbit, which has the constant value $(\mu/a^3)^{\frac{1}{2}}a^2(1-e^2)^{\frac{1}{2}}$, we have

$$\alpha = \frac{r\delta V}{na^2(1-e^2)^{\frac{1}{2}}},$$

where

$$n = (\mu/a^3)^{\frac{1}{2}}.$$

If now we think of a continuously applied perturbing force W per unit mass, perpendicular to the orbital plane, it follows that

$$\frac{\mathrm{d}\Omega}{\mathrm{d}t} = \frac{Wr}{na^2(1-e^2)^{\frac{1}{2}}}\frac{\sin u}{\sin i}, \qquad \ldots\ldots (3.49)$$

$$\frac{\mathrm{d}i}{\mathrm{d}t} = \frac{Wr}{na^2(1-e^2)^{\frac{1}{2}}}\cos u, \qquad \ldots\ldots (3.50)$$

and

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} = -\frac{Wr}{na^2(1-e^2)^{\frac{1}{2}}}\sin u\cot i. \qquad \ldots\ldots (3.51)$$

Perturbations in the orbital plane will have no effect on Ω or i but will produce changes in ω. If R is the component of the perturbing force along the radius vector from the centre and S that along the perpendicular direction in the same plane, in the sense of increasing ϑ, it is found that

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} = -\frac{r}{na^2(1-e^2)^{\frac{1}{2}}}\sin u \cot i W - \frac{(1-e^2)^{\frac{1}{2}}\cos\vartheta}{nae}R + \frac{(1-e^2)^{\frac{1}{2}}}{nae}\left[1+\frac{r}{l}\right]\sin\vartheta S, \quad \ldots . (3.52)$$

a result which may be obtained by elementary arguments depending on the geometry of the ellipse.

The Perturbing Force due to the Polar Flattening

When allowance is made for the polar flattening the gravitational potential of a particle of unit mass in the Earth's field may be written

$$V = \frac{\mu}{r} + \frac{\epsilon_2}{r^3}P_2(\cos\psi) + \frac{\epsilon_4}{r^5}P_4(\cos\psi) \quad \ldots . (3.53)$$

where ψ is the latitude angle (see Fig. 10) and r the distance from the centre. In the notation used by JEFFREYS (1952) we have

$$\mu = gR^2, \ \epsilon_2 = -\frac{2}{3}gR^4J, \ \epsilon_4 = \frac{8}{35}gR^6D, \quad \ldots . (3.54)$$

where R is the equatorial radius of the Earth and g is such that the acceleration due to gravity at the equator as given by (3.53) is equal to the observed apparent gravity at the equator plus the correction for the Earth's rotation. Thus $R = 3963{\cdot}1$ miles, $g = 32{\cdot}146$ ft/sec^2. According to JEFFREYS

$$J = 1{\cdot}637 \times 10^{-3}, \ D = 10{\cdot}6 \times 10^{-6}.$$

To relate ϵ_2 with the difference between polar and equatorial radii we have

$$\epsilon_2 = \mu R^2\left(\alpha - \frac{\tilde{\omega}^2 R}{2g}\right),$$

where

$$\alpha = \Delta R/R, \quad \ldots . (3.55)$$

$R-\Delta R$ being the polar radius, and $\tilde{\omega}$ the Earth's angular velocity. Too much weight should not be given to the value of D, which was chosen so that the Earth's gravitational potential remained constant over a spheroid of equatorial and polar radii R and $R-\Delta R$ respectively.

To avoid complication we shall carry through the analysis neglecting ϵ_4. The components of the force arising from the perturbing potential due to the polar flattening are then given by

$$g_r = \frac{\partial V}{\partial r} = -\frac{\mu}{r^2} - \frac{3\epsilon_2}{r^4}P_2(\cos\psi), \quad \ldots . (3.56)$$

$$g_\psi = \frac{\partial V}{r\partial\psi} = -\frac{3}{2}\frac{\epsilon_2}{r^4}\sin 2\psi. \quad \ldots . (3.57)$$

g_r is the component along the radius vector and g_ψ that tangential to the parallel of latitude in the sense of increasing ψ.

To resolve these into the components W, R and S used in §3, p. 53 we turn to the spherical triangle NQS (Fig. 10) where QS is an arc of the great circle passing through

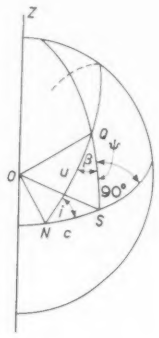


FIG. 10.

the poles. The sides and angles of this triangle are as indicated. Then

$$R = g_r, \quad S = g_\psi \cos\beta, \quad W = g_\psi \sin\beta. \qquad \ldots . (3.58)$$

From the geometry of the right-angled triangle NQS

$$\frac{\sin\psi}{\sin i} = \frac{\sin u}{\sin \pi/2},$$

giving

$$\sin\psi = \sin u \sin i. \qquad \ldots . (3.59)$$

Also

$$\cos u = \cos\psi \cos c,$$
$$\cos u \sin c = \sin u \cos c \cos i,$$

so

$$\sin c = \frac{\sin u \cos i}{\cos\psi}.$$

But

$$\frac{\sin c}{\sin\beta} = \sin u,$$

so

$$\cos\psi \sin\beta = \cos i. \qquad \ldots . (3.60)$$

Finally

$$\cos u \sin\psi = \sin u \cos\psi \cos\beta,$$

so that

$$\sin 2\psi \cos\beta = 2 \cot u \sin^2\psi. \qquad \ldots . (3.61)$$

Using (3.59), (3.60) and (3.61) with (3.56), (3.57) and (3.58), we have

$$W = -\frac{3}{2}\frac{\epsilon_2}{r^4}\sin 2i \sin u, \quad R = \frac{3\epsilon_2}{2r^4}(3\sin^2 i \sin^2 u - 1),$$
$$S = -\frac{3\epsilon_2}{r^4}\sin^2 i \sin 2u. \qquad \ldots . (3.62)$$

Secular Changes of Ω, i *and* ω

To calculate the rate of change of Ω, i and ω per revolution we proceed as in §3, p. 42, writing

$$\dot{\vartheta} = r^{-2}\{\mu a(1-e^2)\}^{\frac{1}{2}}, \qquad \ldots . (3.63)$$

so that, using (3.47) and (3.62)

$$\Delta\Omega = -\frac{3\epsilon_2 \cos i}{\mu a^2(1-e^2)^2}\int_0^{2\pi} \sin^2(\vartheta+\omega)(1+e\cos\vartheta)\mathrm{d}\vartheta,$$
$$= -\frac{3\pi\epsilon_2 \cos i}{\mu a^2(1-e^2)^2}. \qquad \ldots . (3.64)$$

Thus the polar flattening produces a rotation of the orbit about the Earth's axis in a sense opposite to that of revolution of the satellite in its orbit.

There is no secular effect on the inclination for

$$\Delta i = -\frac{3}{4}\frac{\epsilon_2 \sin 2i}{\mu a^2(1-e^2)^2}\int_0^{2\pi} \sin 2(\vartheta+\omega)(1+e\cos\vartheta)\mathrm{d}\vartheta. \qquad \ldots . (3.65)$$
$$= 0$$

The angle ω does undergo a secular change which may be calculated from (3.49), (3.62) and we find

$$\Delta\omega = \frac{3\pi\epsilon_2}{2\mu a^2(1-e^2)^2}(4-5\sin^2 i). \qquad \ldots . (3.66)$$

It follows that the polar flattening causes a rotation of the orbit in its own plane except when $\sin^2 i = 4/5$, $i = 64°5$. For inclinations less than this the rotation is in the same sense as the revolution of the satellite, while for greater inclinations it is retrograde.

A number of authors (SPITZER 1950, BLITZER *et al.*, 1956, 1957a, 1957b, ROBERSON, 1957, OKHOTSIMSKII, ENEEV and TARATYNOVA, 1957, YATSUNSKII, 1957, TARATYNOVA, 1957, and KING-HELE, 1958) have discussed the secular perturbations. KING-

HELE and ROBERSON give formulae which include higher order terms. Thus KING-HELE gives

$$\Delta\Omega = -\frac{3\pi\epsilon_2 \cos i}{\mu a^2(1-e^2)^2}\left[1 - \frac{JR^2}{a^2(1-e^2)^2}\left(\frac{19}{12}\sin^2 i - 1\right) - \frac{3DR^2(7\sin^2 i - 4)}{14Ja^2(1-e^2)^2}\right]. \quad \ldots . (3.67)$$

The latter author also shows that the radial distance of the satellite from the Earth's centre undergoes a periodic change δr from the value it would have for an unperturbed elliptic orbit. δr is given by

$$\delta r = -\frac{JR^2}{6a(1-e^2)}\sin^2 i \cos 2u. \quad \ldots . (3.68)$$

Finally, the period of revolution becomes

$$\tau = \tau_0\left\{1 - \frac{JR^2}{a^2(1-e^2)^2}\frac{7\cos^2 i - 1}{4} + O(J^2)\right\}, \quad \ldots . (3.69)$$

where τ_0 is the period for the spherical Earth.

The Effect of a Third Harmonic in the Gravitational Potential

KING-HELE (reported by CORNFORD 1958) has calculated the contribution to the secular variation of Ω due to the presence in the gravitational potential (3.53) of an additional term

$$\frac{\epsilon_3}{r^4}P_3(\cos\psi).$$

This adds a term

$$\frac{-3\epsilon_3 e}{8JR^3a(1-e^2)}\sin\omega \operatorname{cosec} i \ (15\sin^2 i - 4) \quad \ldots . (3.70)$$

within the bracket of (3.67).

Combined Effect of Air Drag and Polar Flattening

TARATYNOVA (1957) has carried out a detailed calculation of the changes in the orbit elements with time for a particular case, allowing both for air drag and the polar flattening. This involved the solution of the equations (3.47), (3.48) and (3.49), together with those for the semi-latus rectum l and eccentricity e, i.e.:

$$\frac{\mathrm{d}l}{\mathrm{d}t} = 2r(l/\mu)^{\frac{1}{2}}S,$$

$$\frac{\mathrm{d}e}{\mathrm{d}t} = (l/\mu)^{\frac{1}{2}}\sin\vartheta R + (l/\mu)^{\frac{1}{2}}\{1 + (r/l)(\cos\vartheta + e)\}S,$$

and that representing the change of the true anomaly ϑ with the time:

$$r^2\left(\frac{\mathrm{d}\vartheta}{\mathrm{d}t} + \frac{\mathrm{d}\omega}{\mathrm{d}t} + \cos i\frac{\mathrm{d}\Omega}{\mathrm{d}t}\right) = (l\mu)^{\frac{1}{2}}.$$

The calculation was carried out with an electronic computer. A spherical satellite of mass 10 kg and radius 25 cm was assumed, together with the initial orbital elements h(apogee) = 1284 km, h(perigee) = 320 km, $i = 45°$, $\omega = 90°$, $\Omega = 129°$. The drag coefficient was taken as 2 and the air density as given by (3.48) with the constants of Table 2.

Figure 11 illustrates the variation of the different elements with time over the

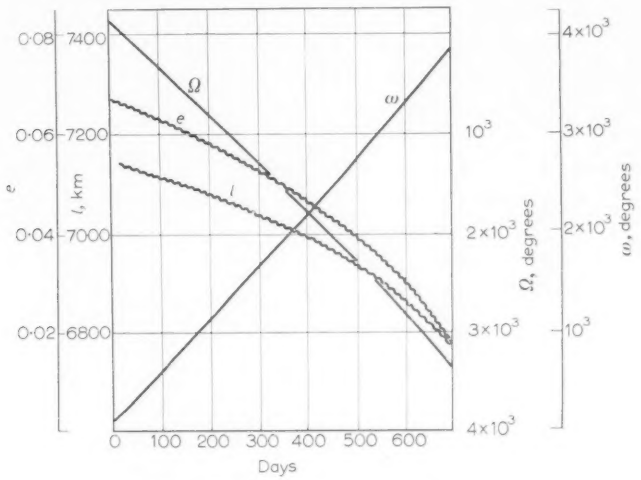


FIG. 11. Variation of Ω, ω, e and l with time for a particular but typical case of a satellite orbit, calculated accurately by numerical computation.

first 700 days. The secular changes of Ω and ω due to polar flattening and of l and e due to the air drag are clear. It will be seen that there is also present a periodic variation of l and e of period about 36 days. This corresponds to the time for ω to make a complete revolution so that the perigee height of the satellite goes through a cycle because of the polar flattening.

Derivation of New Data about the Figure of the Earth

The application of the formulae of this section to the analysis of the motions of the satellites already launched has provided new information about the figure of the Earth, even at this early stage. An account of this work will be given in §13, p. 99 *et seq.*

Future work in this direction will include consideration of periodic as well as secular effects but this will require increased accuracy in determination of position and time data for satellites.

4. METHODS FOR TRACKING SATELLITES

Three methods are in principle available for observing the path of a satellite. First of all there is the optical method which may be either visual, photographic or photoelectric in character. Special problems arise in utilizing this method because of the high speed of a satellite and it is very difficult to use it if information is not already available to indicate the part of the sky towards which the optical instru-

ment should be pointed at a particular time. On the other hand if such information is available it is potentially possible to determine the position of the satellite at a particular time with greater precision than by any other method.

The second method depends on the presence of a radio transmitter within the satellite. Radio signals may be received by suitable ground equipment, which does not have to be oriented in any particular direction, and the approximate angular coordinates of the satellite at different times may be determined to a useful accuracy. This method is therefore most valuable for location of the satellite soon after launching and for provision of predictions of satellite position so that observers with precision optical equipment can direct their instruments suitably. On the other hand the intrinsic accuracy of the radio method is not as high as for the optical method.

A further method, which in some ways is intermediate between the other two, is the use of radar. This has the advantage of requiring no radiation from the satellite itself so can be used when the satellite transmitter has ceased operation and when it is eclipsed by the Earth's shadow. It is possible at some sacrifice in accuracy to use a radar beam width which makes detection possible even when only very crude predictions are available. It is not capable of accuracy comparable with the optical method but there are circumstances when it is the only method which may be used.

We shall not attempt here to go into detail about the various techniques which are being employed for accurate position determinations by these different methods but shall give some general account of the principles involved and the performance expected of some of the more important equipment.

Radio Tracking

An important choice which must be made at the outset is that of the frequency of the signal to be radiated from the satellite beacon. If it were necessary to rely on radio tracking alone the choice would be guided by the need to minimize errors due to the effect of the ionosphere below the satellite. This will cause refraction and also modify the Doppler shift due to the satellite's motion (see §11, p. 82). To minimize these effects a high frequency should be employed. On the other hand, by so doing, all prospect of obtaining much information about many features of the ionosphere is lost. When one allows for the fact that radio tracking of itself is never accurate enough to provide really precise orbit data and that optical methods may be used for this once approximate data are available, a compromise choice becomes possible.

The frequency of 108 Mc/sec chosen for the first American satellites comes into the first category while the Russian choice of 20 and 40 Mc/sec falls into the second. With these latter two frequencies much data of interest for the ionosphere can be obtained while at the same time the higher frequency gives results of sufficient accuracy for acquirement and prediction.

The apparent angular position of a satellite which is transmitting radio signals may be determined by use of a radio-interferometer of a type similar to that employed in radio-astronomy. Essentially the method depends on the accurate measurement of the phase differences between signals received at suitably placed aerials whose separations are very accurately known. Basically three aerials not in a straight line would suffice to determine two angular coordinates but for the need to resolve ambiguities in phase. With two aerials only, the phase difference from which a single angular coordinate is derived is ambiguous by an integral multiple of π. This is

removed if a third aerial in the same straight line is used, provided the separations of the aerials are not in simple integral ratios to each other. Two American systems of phase comparison have been developed, the Minitrack and the Microlock, both of which give precise apparent angular position of a transmitting satellite. The adjective "apparent" allows for the fact that there will be some disturbance due to the ionosphere, apart from any inaccuracy in the receiving equipment. Even at the relatively high frequency of 108 Mc/sec at which these systems operate the ionospheric errors are considered to be greater than those arising from other sources. For a satellite at a zenith angle of 45° the ionosphere introduces an error in angular position of about $\frac{1}{4}°$, but at smaller zenith angles it is much less, being only about 80″ (see §11, p. 82). On the other hand these errors are predictable if the ionospheric structure is known—in practice this means that they may be reduced to between one third and one quarter of the values mentioned.

Radio-interferometers operating at 20 and 40 Mc/sec were improvised for tracking the Russian satellites and gave valuable data even though they were somewhat less accurate than the more elaborate Minitrack and Microlock systems. At 40 Mc/sec the ionospheric corrections are about 6 times larger than for 108 Mc/sec (see p. 82).

A second method of radio tracking depends on observation of the Doppler effect exhibited by the signal from the moving satellite beacon. If f_0 is the frequency of the transmitter, the observed frequency f when the satellite is at a distance ρ from the receiving station is given by

$$f - f_0 = \frac{f_0}{c}\frac{\mathrm{d}\rho}{\mathrm{d}t}. \qquad \ldots . (4.1)$$

The time t_0 of closest approach ($\rho = \rho_0$) will be such that $\mathrm{d}\rho/\mathrm{d}t = 0, f = f_0$. At this point the rate of change of frequency, given by $f_0\ddot{\rho}/c$, will be a maximum. Since $\ddot{\rho}$ will be given by V^2/ρ_0, where V is the satellite speed,

$$\dot{f}_{\max} = \frac{f_0}{c}\frac{V^2}{\rho_0}. \qquad \ldots . (4.2)$$

The variation of the Doppler frequency with time is thus as indicated in Fig. 12. Analysis of a frequency-time record is first directed towards the determination of t_0 and hence of f_0. From (4.2) V^2/ρ_0 may then be obtained. Use may be made of the observations at other times to determine V and ρ_0 separately.

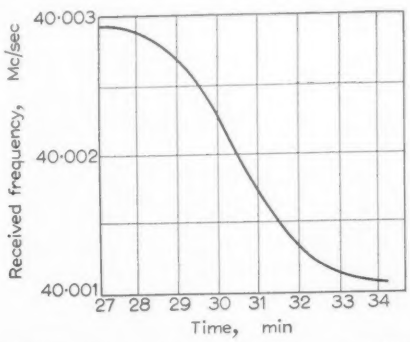


FIG. 12. A typical time variation of received frequency from a satellite radio beacon. This is an observed curve for Sputnik I at a slant range of 650 km, taken by the staff of the Mullard Radio Astronomy Observatory, Cambridge, on 1957 October 9.

Thus, if time τ is measured from the instant of closest approach we have, if curvature of the orbit and of the Earth is neglected, that

$$\rho^2 = \rho_0^2 + V^2\tau^2,$$

or

$$\frac{\tau^2}{\dot{\rho}^2} = \frac{\rho_0^2}{V^4} + \frac{\tau^2}{V^2}. \qquad \ldots . (4.3)$$

Using values of $\dot{\rho}$ derived from the frequency-time curve a plot of $\tau^2/\dot{\rho}^2$ against τ^2 should give a straight line from which V and ρ_0 can both be obtained approximately. Allowance may then be made for the effect of curvature to give more accurate results.

The presence of the ionosphere introduces a modification of the Doppler effect which is discussed in some detail in §11. For a natural frequency of 108 Mc/sec the frequency shift due to the ionosphere, for a ray incident at 45°, is only about 1 cycle/sec but is about 6 times greater at 40 Mc/sec.

Given two interferometers or Doppler recording stations at different locations it is possible to determine all three coordinates fixing the position of the satellite at the times of observation. This can also be done if interferometer and Doppler recordings can be made at the one site.

Optical Tracking

The accuracy of observation required to derive orbit data of such precision as to yield new results about the figure of the Earth and for geodesy is high. For these purposes resort will ultimately have to be made to the study of periodic as well as secular perturbations. This will require measurements of angular positions to about 0·001° and of time to about 0·001 sec. Less accurate data are required for initial analysis of secular effects (see §13).

It is quite clear that the only hope of attaining such precision is by optical methods. Even so, the problem is very different from the usual astronomical one because of the high speed of a satellite which moves with an angular speed of between 1000 and 5000 sec of arc per sec. If time is measured to 0·001 sec the possible accuracy in angular position is restricted to between 1 and 5 sec of arc which is only just within the required range. This precludes the use of large astronomical telescopes of long focal lengths and turns attention towards instruments of the Schmidt type as used for meteor studies.

The Baker-Nunn camera, specially designed for satellite tracking (Whipple and Hynek, 1956), is a wide angle camera of this type in which the optical system consists of a 31 in. spherical mirror and three 20 in. correcting lenses with photographic strip film stretched over a curved surface at the focus of the mirror. By means of a control system the camera can be set to face any part of the sky and follow automatically the predicted path of the satellite. On each strip of film two pictures are taken, one while the camera is fixed on and following the satellite, and the other while the camera is fixed on and moving with the background stars, thus providing reference positions against which the satellite may be located. A crystal clock operates a set of time marks on the film correct to 0·001 sec. With this system and ordinary

fast photographic emulsions it is possible to track objects considerably fainter than the 6th magnitude moving at an angular speed of 1° per sec.

Kinetheodolites of standard design may be used to determine position to about 0·01°.

A photoelectric method of precision tracking has been developed by WILLMORE (1958). A system of M-shaped slits is placed in the focal plane of a lens. The passage of the image of a satellite (when sunlit) across the slits produces a series of pulses in a photomultiplier located behind them. With this system a long focal length may be used, so simplifying the design of the optical system. This arrangement seems to be capable of high accuracy.

Radar Tracking

The accuracy of determination of angular position by radar is limited by the beam width of the transmitting aerial which is usually of the order of a few degrees. Thus, for the tracking of the first two Russian satellites by the aerial at Jodrell Bank the beam width was about 8°. It was not possible to use a finer beam as in many cases the radar method was being employed to acquire the satellite. The 45 ft radio telescope at the Royal Radar Establishment, Malvern, employs a beam width of $\frac{1}{2}$° but it could not be used unless quite accurate predictions of the satellite track were available.

The accuracy of range determination is fixed by the length of the transmitted pulse. This may be as short as, say, 5 μsec giving a precision of about 1·5 km in range. This is a useful accuracy for the determination of perigee heights as the error is considerably smaller than the atmospheric scale height which is likely to be 20 km or more at satellite altitudes. Radar range data of this accuracy could therefore be used in the determination of air density (see §13).

The chief value of radar tracking lies in its applicability when the radio transmitters have failed or when the satellite object, such as a final stage rocket case, carries no radio transmitter. This was clearly shown in the case of the first Russian satellites and is often likely to apply during the particularly interesting last stages of the life of a satellite.

Utilization of Relatively Inaccurate Observations

Amateur observers may play an important part in the initial stages when acquirement is still incomplete, but they can play an even more vital role in the last stages when the orbit is changing rapidly and the ultimate fate of the object is difficult to observe. So little is known about the sequence of events as a satellite plunges into the atmosphere that considerable emphasis must be placed on any observations which throw light on it.

In other aspects of tracking the accuracy requirements are so high that very little weight can be given, even for prediction purposes, to observations in which an accuracy as high as 0·1° in position and 0·1 sec in time cannot be assumed.

Radio Reflections from Satellite Ionization Clouds

KRAUS (1958) has reported observations which suggest that a satellite produces a degree of excess ionization sufficient to produce reflection of radio waves of 20 Mc/sec frequency. He made observations of the intensity of the signal on this frequency,

radiated in the WWV transmissions from Washington, received at Columbus, Ohio, 300 miles away. During daytime the ionosphere is sufficiently dense to produce reflection and hence a strong signal is received. At night, on the other hand, the weakened ionosphere cannot produce reflection and no signal is normally received. Short bursts, however, are observed for periods of a few seconds due to sporadic excess ionization by meteors. KRAUS found that bursts of longer duration (a few minutes) were received close to the times of passage of Sputnik I and Sputnik II. Assuming that the correlation is a real one he was even able to follow the disintegration of Sputnik I into fragments on entering the atmosphere.

KRAUS suggests that these signal bursts are due to reflection from excess ionization produced by the satellite as it passes through the air. On many occasions (KRAUS and ALBUS, 1958) the signals received at 20 Mc/sec from the satellite radio beacon were observed to be of very irregular amplitude while the satellite was approaching but changed to a regular form as it passed. This KRAUS interprets as due to the passage of signals from the approaching satellite through the satellite-produced ionization which is concentrated in a conical region ahead of the satellite.

The difficulty about this explanation is that of understanding how a satellite can produce an appreciable ionization. An atmospheric gas molecule colliding with the satellite cannot acquire an energy greater than 10 eV, which is too low for it to produce impact ionization in the air. It is also unlikely that appreciable secondary emission of charged particles resulting from bombardment of the satellite surface could be produced in such low energy impacts. One might suppose as an alternative source of ionization that the satellite charges up to a negative potential of many tens of volts relative to the surrounding air so that the positive charges already present produce further ionization as they are accelerated towards the satellite. This might possibly be important in daytime (see §3, p. 40) but is most unlikely to occur at night during which period the satellite potential is not likely to differ from that of the surrounding air by more than a few volts. The whole question needs much further exploration.

5. Temperature of a Satellite and its Control

The electrical, including electronic, equipment in a satellite will only operate efficiently if the temperature remains between fairly narrow limits. It is important to employ power sources in a satellite which give a high power output per pound of mass. Two types of cell which are most efficient in this sense are the mercury alkaline and zinc silver-oxide cells. The efficiency of the former begins to fall as the temperature decreases below about 15°C and is effectively zero at 0°C. With the zinc silver-oxide cell the lower limits of useful operation may be extended by perhaps 15°C. Other cells have efficiencies which vary more slowly with temperature at low temperatures but the maximum efficiencies obtainable with them are considerably lower.

Near the upper temperature limit the efficiency falls much more slowly with temperature increase but is beginning to pass the maximum as a temperature of 60°C is exceeded. An upper limit of this magnitude must be imposed in any case in order that transistors may be used in the electronic circuits.

In the design of the American Vanguard satellites, which are spherical in shape, the instruments were contained within a central package which could be thermally

insulated to a high degree from the shell. This may be secured by supporting the package with materials of low thermal conductivity and by gold-plating the inside of the shell and outside of the instrument container to reduce radiation transfer. Once the heat transfer properties between shell and container have been measured and the mean amount of power dissipated in the instrument container during flight is known it is possible to assign limits within which the temperature of the satellite shell must lie in order that the instruments will function effectively.

The temperature of the shell is determined almost completely by a balance between the radiant energy received from the Sun and Earth and that radiated by the shell. Opportunity is afforded for controlling the temperature by choosing the shell surface to have different absorbing power towards solar and infra-red radiation.

If a_s is the absorptivity of the shell material towards solar radiation and a_e towards infra-red radiation in the temperature range 0–60°C the absolute temperature T of the shell is determined by the equation

$$R_s a_s + R_r a_s + R_e a_e - A\sigma a_e T^4 = ms\frac{\mathrm{d}T}{\mathrm{d}t}. \qquad \ldots\ldots (5.1)$$

R_s is the solar radiant energy incident directly on the shell, R_r that reflected from the Earth. R_e is the infra-red radiation from the Earth. A is the surface area, m the mass and s the specific heat of the shell, and σ is Stefan's constant.

For a given position of the satellite in relation to the Sun it is possible to calculate R_s with considerable accuracy but R_r and R_e are more difficult to predict. This is because the Earth's albedo, which determines R_r and the intensity of infra-red emission from the Earth's atmosphere, can be expected to vary in an irregular manner depending on cloud cover, etc. In fact a study of satellite shell temperature is likely to provide valuable new information on these matters. In any case the shell temperature will vary in a quasi-periodic fashion between the sunlit and dark phases of the satellite's motion. Figure 13 illustrates the variation for a particular case worked out by DRUMMETER and SCHACH (1957). This is obtained by assuming the Earth's albedo to have the constant value 0·36.

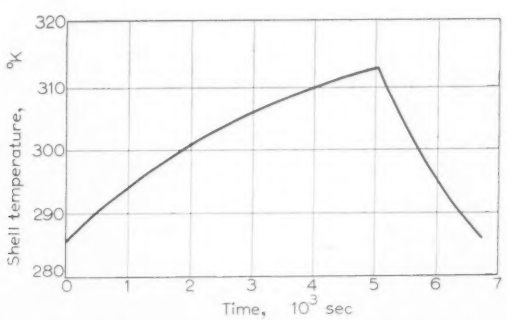


FIG. 13. Variation of satellite shell temperature with time, calculated for an orbit with perigee and apogee heights of 200 and 1500 miles respectively, taking $a_s = 0{\cdot}20$, $a_e = 0{\cdot}164$ and supposing that for 75 per cent of the period the satellite is sunlit.

The temperature attained by the instruments within the container will be determined by a suitable mean of the temperature of the shell throughout the orbital

circulation. To a good approximation this may be obtained from (5.1) by putting the right hand side equal to 0 and substituting average values over the orbit for R_s, R_r and R_e. Thus, if T_m is the required mean temperature, we have

$$A\sigma T_m^4 = \bar{R}_e + (\bar{R}_s + \bar{R}_r)\frac{a_s}{a_e}. \qquad \ldots . (5.2)$$

For a spherical satellite R_e, R_s and R_r are proportional to A so T_m is independent of A and is determined for a given orbit simply by the ratio a_s/a_e. Figure 14 illustrates the variation of T_m with a_s/a_e for a typical case.

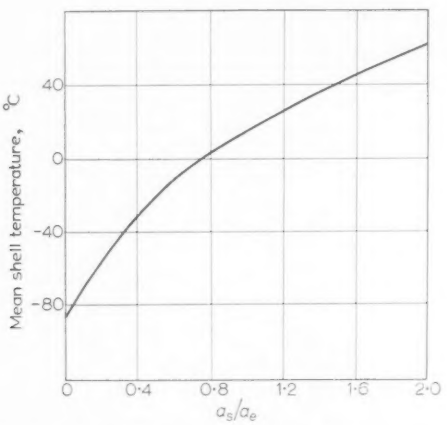


FIG. 14. Variation of mean satellite shell temperature (averaged over a revolution) with the ratio a_s/a_e. The perigee and apogee heights are 200 and 1500 miles respectively, 75 per cent of the period is sunlit, the Earth's albedo is taken to be 0·5 and the season as winter.

To confine the temperature of the instruments within the required limits for a satellite of the Vanguard type, in which the internal power dissipation is about 400 mW, the ratio a_s/a_e should be 1·15. It is not easy to reconcile this with the requirement of high visibility necessary for optical tracking. Thus a polished metal surface, suitable for the latter purpose, has a ratio a_s/a_e of about 5 and would lead to unacceptably high internal temperatures. A way out of this difficulty, which is being employed in the Vanguard satellite, is to coat the highly reflecting surface with a thin layer of a dielectric which is transparent to visible radiation but opaque in the infra-red, so that, while a_s is unaffected, a_e can be increased to the required extent.

The measurement of shell temperature while a satellite is in its orbit will be important, not only to provide design data for the instrument containers of future satellites but also for its intrinsic interest.

Thus, it should be possible from a study of its variation with phase in the orbit to determine whether the ratio a_s/a_e is constant and if this should prove to be so, information about the Earth's albedo and radiation intensity can be derived. On the other hand, as the satellite will be exposed to meteor bombardment as well as irradiation by ultra-violet light, the optical properties of the surface of the shell may change with time. This should be detectable from the temperature measurements and correlation with results of other measurements on shell erosion, etc., could be sought.

Shell and interior temperatures may be measured by means of thermistors or other resistance gauges. An auxiliary method which has been used in Explorer and in Vanguard satellites is to monitor the frequency of a radio transmitter in the satellite which is temperature sensitive and has been calibrated before launching. Results have demonstrated the effectiveness of the temperature controls.

6. Meteorological Studies with a Satellite

The value for meteorology of continuously available information about such matters as the distribution of cloud over the globe should be very considerable, as at present only about 5 per cent of the Earth's atmosphere is covered by observations. Ultimately one can expect that observations of cloud cover will be made by television with telemetry to ground of the pictures. In earlier satellites resort must be made to much simpler systems. One of these which it was proposed to employ in Vanguard satellites consists of a rudimentary scanning device in which three photocells in the satellite can receive radiation incident in narrow solid angles along directions fixed with respect to the axis of spin. As the satellite rotates and revolves in its orbit the cells sweep out different parts of the Earth's atmosphere. From the telemetred signals received from each cell it will be possible to reconstruct the orientation of the satellite at each point of the orbit and hence determine the area of the Earth's atmosphere from which the radiation is received in each cell. The degree of cloud cover over the area can be estimated from the fact that the albedo of a cloudy surface is very much higher (about 80 per cent) than that of land (15–20 per cent) or sea (5 per cent). In the earliest experiments the areal resolution likely to be achieved will not be worse than 1000 mile^2 and will be very much better under most conditions. This should be sufficient to obtain information about the structure of large scale disturbances.

Another experiment proposed is aimed especially at studying the radiation balance of the Earth by making direct measurements of the intensities of incoming and outgoing radiations from the Earth's surface as the satellite revolves.

Although a good deal of knowledge is now available about the annual radiation balance the situation concerning the balance over shorter periods is much less satisfactory due to lack of sufficient observations. It is proposed to determine separately the flux of direct solar radiation (r_s) of solar radiation reflected from the Earth's atmosphere (r_a) and of the infra-red radiation (r_e) from the Earth, by the following technique.

The radio antennae of the satellite are used to mount four small spheres. Two of these are about $1\frac{1}{4}$ in. diameter coated with different materials and exposed to the solar radiation while the other two are smaller and shaded from the Sun. The data recorded and telemetred to ground are the temperatures of the four spheres as measured by thermistors.

If, at any point of the satellite's path, the Earth subtends a solid angle β at one of the small spheres the equilibrium temperature of the sphere is given by the radiation balance equation (5.2) where, in terms of the radiant fluxes r_s, r_a or r_e we write

$$\pi r_s = \bar{R}_s, \ \beta r_a = \bar{R}_a, \ \beta r_e = \bar{R}_e.$$

Three equations from which to determine the three unknown radiant fluxes may be obtained from the measured temperatures for three spheres coated with materials

with markedly different values of a_s/a_e. Thus for black paint $a_s/a_e = 1$ and for white paint about 0·05. A third coating developed by TABOR (1956) gives a ratio of 9. By measuring the equilibrium temperature for a black-painted sphere and a white-painted sphere exposed to the Sun, and also that of a Tabor-coated sphere shielded from direct solar radiation, it is possible to obtain r_s, r_a and r_e with good accuracy. SUOMI (1957) has discussed the application of this technique with satellites whose orbital inclinations are near 40° and shown that information could be obtained of much value for the detailed day to day radiation balance, particularly over the tropical belt for which there is a net input of radiation.

7. The Measurement of Atmospheric Pressure and Density

A knowledge of pressure, density and temperature as a function of height is basic for atmospheric physics. Even at altitudes between 60 and 100 km much remains to be done before accurate information is available, particularly about variations with time and place. Before the launching of the first artificial satellites only two density measurements had been made with rocket-borne pressure gauges above 100 km, one at 160 km and the other at 219 km, both at White Sands, New Mexico. Since that time more extensive observations have been made up to 200 km at Fort Churchill (LAGOW, HOROWITZ and AINSWORTH 1958). These observations indicate a strong sensitivity of atmospheric structure at altitudes above 100 km to time of day, latitude and season. The need for much more extensive observations, particularly at altitudes above 200 km, is obvious. In view of the increased variability indicated from the rocket observations, the opportunities afforded for systematic observation by the availability of artificial satellites are of special value.

The possibility of obtaining information about the density of the atmosphere at heights near the perigee of a satellite orbit by observing the variation of orbital period with time has already been discussed in §3, p. 49. While most useful data are being obtained in this way (see §13, p. 97), the method is limited by the fact that it is essentially statistical in character, depending on averaging over a great number of orbital revolutions. Rapid variations with position and time would thus be averaged out. It is therefore desirable to seek additional methods which yield results over a small portion of an orbit.

A number of suggestions have been made for doing this. Some depend essentially on measurements of the air drag over small parts of the orbit and some on the use of ionization gauges of one type or other to measure pressure or density directly.

Mechanical Methods

Among the first type is the suggestion by SPITZER (1957) of using change in orientation or precession of a suitably shaped satellite to measure the drag. Thus if there is attached to the main, say spherical, satellite a second one of much smaller mass but similar cross-sectional area so the line of centres is nearly perpendicular to the direction of motion, a torque will be exerted about the centre of mass of the system by the air drag. If the "dumb-bell" is launched without axial rotation the torque will produce rotation about an axis perpendicular to that of the "dumb-bell" and to the direction of motion. Alternatively, if the "dumb-bell" is launched with a rotation about an axis of symmetry the torque will produce a rotation about an axis in the direction of motion. In either case the change of orientation could be observed

to within about 3° by using photoelectric scanners in the satellite (cf. §6). SPITZER shows that with practicable magnitudes for the system it should be possible to determine average values of air density over a time interval as short as 20 min at altitudes up to 400 km or so.

A second suggestion by JONES and BARTMAN (1957) is to extend a technique already utilized with some success in rocket experiments. This is to eject at the required altitude an inflatable envelope containing a sensitive accelerometer and a radio-beacon for tracking. Soon after ejection the envelope is inflated to a spherical surface of 5 ft diameter. This provides a body of high cross-section to mass ratio. The weight of the equipment need not exceed 5 lb so that its motion is very sensitive to air drag. A similar principle could be used by launching the envelope and equipment as a satellite. To measure the air drag at altitudes of 300–400 km it would be necessary to increase the sensitivity of the accelerometer by from 10 to 100 times over that used in the rocket experiments. This does not seem to be impracticable.

An unsuccessful attempt to launch a large inflatable satellite of this kind was made in the U.S.A. during 1958.

Electrical Methods

Although the useful operation of an ionization gauge is limited to pressures above 10^{-8} mm Hg it is possible to extend this range somewhat by suitable modifications. The limitation arises from the occurrence of photoelectric emission from the anode due to bombardment by soft X-rays generated by the slowing-down of electrons when they collide with the grid. By reducing the surface area of the anode to the minimum and making the grid of a material of low atomic number this source of background current may be decreased so that the limiting pressure may be lowered by a factor of 100 or so. For use in a satellite there is the essential practical requirement that the cathode emission be automatically stabilized. Apart from this there still remain difficulties of interpretation depending on atmospheric composition.

The gauge must be connected to the air through some orifice or tube and the relation between the density of the ambient air and the pressure as recorded by the gauge depends on the geometry of the system and the mean molecular weight of the air. Thus, suppose the gauge is exposed to the atmosphere through a plane orifice inclined at an angle θ to the direction of motion of the satellite. If the atmospheric gas were composed of particles of mass m, in concentration N per cm^3 the pressure P_1 recorded by the gauge would be given by

$$N = \frac{P_1}{(2\pi mkT_1)^{\frac{1}{2}}} \frac{1}{U \sin\theta}, \qquad \ldots . (7.1)$$

where T is the temperature of the gas in the gauge and U is the speed of the satellite. For a gas of mixed composition a suitable mean must be used for m. Considerable uncertainty would therefore be attached to the interpretation of the gauge readings as, particularly at very high altitudes, the atmospheric composition is not well known. On the other hand, our knowledge of the density at such altitudes is so slight that in the first instance the uncertainties introduced by the mean molecular mass would be tolerable.

Nevertheless, at some stage, it would be desirable to use a technique in which these uncertainties may be eliminated. The omegatron offers such a possibility. This

device is an ionization gauge which uses the principle of the cyclotron to measure the contribution to the ionization arriving from a specific ion. In principle, with suitable calibration, it may be used to measure the partial pressures of the different components in a gas mixture so that the formula (7.1) could be applied to determine the concentration of each.

Figure 15 illustrates the way in which the omegatron operates. The gas under investigation is contained in a space pervaded by a uniform magnetic field H. Ionization is produced by an electron beam B in the direction of the field. An alternating electric

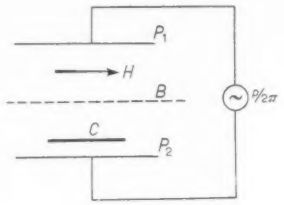


Fig. 15.

field of magnitude E and frequency $p/2\pi$ is applied to the plates P_1, P_2 in a direction perpendicular to the field. In general the slow ions produced are confined within the region of the beam by the magnetic field but, if there are ions present whose mass m is such that

$$p = eH/m,$$

they will be accelerated as in a cyclotron and describe spiral orbits of increasing diameter until reaching a collector C. In this way, either by varying H or p, it is possible to measure the ionization currents arising from each kind of ion produced.

For application in a satellite advantage can be taken of the fact that nitrogen and oxygen will certainly be the dominant constituents so that five separate gauges set to deal with N_2^+, O_2^+, N^+ and O^+ ions respectively would obviate the necessity for scanning a range of magnetic field or frequency. The solar radiation could be used to provide a photo-electric source for the electron beam and comparison of data obtained when the satellite is illuminated and when it is dark could distinguish between ions present normally in the atmosphere and those produced as impact ionization by the electron beam.

In the early satellites it is unlikely that such complex apparatus will be carried, but as development proceeds there is little doubt that it will become practicable and fruitful.

8. Study of Short-Wave Radiations from the Sun and Other Sources

The introduction of techniques for measuring the intensities of different ultraviolet and X-radiations at heights above 60 km by rocket-borne equipment has provided data of the greatest importance for the understanding of the various processes which determine the nature and behaviour of the ionosphere. Among these radiations the Lyman α line at 1216 Å is of especial interest. This is emitted quite strongly by the Sun under normal conditions and is able to penetrate quite deeply into the atmosphere (down to 80 km) due to a "window" in the absorption by O_2.

It is considered likely that it is responsible for the normal day-time sub-E ionization, rather loosely described as the D-layer. The quantum energy of the radiation, 10eV, while insufficient to ionize the main atmospheric constituents, will ionize nitric oxide the ionization potential of which is as low as 9·2 eV (as we shall see below use is made of this in the design of Lyman α detectors). While we have no evidence about the nitric oxide concentration below the E layer it does not have to be unreasonably large in order that the low ion concentration of the normal D-layer should be produced.

When a solar flare occurs, there is often a breakdown in radio communications over the sunlit hemisphere. This radio fadeout is due to a sudden increase in ionization below the E layer. At these levels the gas pressure is so high that an electron set into oscillation by the oscillating electromagnetic field of a radiowave is likely to lose its surplus energy in collision before it has time to radiate it again. When the electron concentration is high enough, radiowaves entering the region lose all their energy in this way in heating the gas and hence never return to ground. The source of the enhanced ionization is still uncertain although it is possibly due to increased intensity of emission of Lyman α radiation in association with the flare.

There is no positive evidence in support of this and certain difficulties arise when it is considered in detail. It is difficult to carry out rocket launchings to the tight schedule necessary to determine whether there is an enhanced emission of Lyman α radiation by the Sun when a flare occurs. Some experiments on these lines have indeed been carried out as part of the U.S. programme for the IGY (FRIEDMAN *et al.*, 1958) and indicate at present no marked change in the Lyman α intensity shortly after a flare. However, some delay is inevitable between the observation of a flare and the rocket launching. It is therefore of much interest to make continuous recording of the Lyman α intensity from the Sun by means of suitable instruments within a satellite. The importance of such work is indicated by the wide variation in the intensity encountered in different rocket flights even when there is no clear correlation with solar activity. Values ranging from 0·1 to 4 or more erg cm^{-2} sec^{-1} have been obtained. It is not clear to what extent these variations are instrumental. This point could also be settled by synoptic observations from satellites.

The apparatus required is essentially the same as that used by CHUBB, FRIEDMAN and KUPPERIAN (1957a) for flights in Aerobee rockets. It consists of an ionization chamber filled with nitric oxide. Ionization can only be produced by radiation with wavelengths shorter than 1340 Å. Addition of a lithium fluoride window introduces a cut-off for radiation shorter than 1100 Å. This restricts the sensitivity of the chamber to a narrow wavelength range about the Lyman α line.

The comparative simplicity of this equipment makes it suitable for use in quite early satellites so that information about the solar emission of the Lyman α line should be available quite soon. CHUBB, FRIEDMAN and KUPPERIAN have suggested a further experiment of great interest as indicating the possibilities which will be afforded for astrophysics.

The Lyman α radiation from the Sun will produce fluorescent emission from the surrounding interplanetary atomic hydrogen. In addition to this back-scattered radiation there will also be recombination radiation arising from radiative capture of interplanetary electrons by protons. The total intensity of this Lyman α radiation can be measured in daytime by a suitable detector pointing away from the Sun. The intensity of the scattered radiation is proportional to the concentration N_H of neutral

hydrogen atoms and that of the recombination radiation to $(N_{H^+})^2$ where N_{H^+} is the proton concentration. To disentangle the two contributions, observations of the change in intensity during a solar disturbance should suffice. The scattered radiation should change far more rapidly than that due to recombination.

To implement these ideas it would be necessary to include apparatus of high sensitivity to measure the Lyman α intensity emitted in the direction opposite to the Sun, as well as of low sensitivity to monitor the intensity emitted by the Sun. It is possible to do this by using photon counters. These are similar in principle to Geiger counters, though triggered by the selected ultra-violet radiation instead of by fast particles. For both sensitive and insensitive counters a lithium fluoride window excludes short-wave radiation from entering. Triggering by photoelectric emission from the cathode due to long-wave-length radiation may be prevented by including an electro-negative gas as quencher. The sensitive counters would have nitric oxide as the main gas while the insensitive ones need only include chlorine.

Chubb, Friedman and Kupperian (1957b) have examined the quantitative possibilities and find that it should be practicable to detect a concentration of hydrogen atoms as low as $10^{-4}/cm^3$ and of protons down to $10/cm^3$. In addition it should be possible, by observations at night, to obtain information about the Lyman α radiation from interstellar space.

The practicability of this programme has been confirmed by the rocket observations carried out in 1957 by Kupperian and Friedman and their associates (Kupperian 1958a, b). An Aerobee rocket was fitted with an ultraviolet telescope. This consisted of bundles of fine tubes through which the sky could be viewed by sensitive photon counters, the angular resolution being 3°. The angular motion of the rocket ensured that a wide scan of the sky was obtained during the flight. By use of suitable absorbers certain of the detectors were sensitive to the wavelength range 1050 Å to 1225 Å, including Lyman α, and the other to the range 1225 Å to 1350 Å. The rocket was launched at White Sands at 9.56 p.m. local time.

At an altitude of 75 km a strong glow in the former band was observed from the sky above the rocket. When the altitude reached 85 km an additional glow from the atmosphere below the rocket became apparent. Both the inward and outward fluxes increased up to 120 km altitude and then remained constant up to the highest altitude attained (140 km). The intensity of the radiation is a minimum in the anti-solar direction. It is reasonable to assume that the main source of the glow is Lyman α radiation from the Sun. Scattering from the neutral hydrogen in interplanetary space provides the flux from the sky while back scattering from hydrogen in the atmosphere is responsible for the flux from below.

Further evidence in support of this viewpoint is afforded from the observations made in the longer wavelength range. These showed the existence of a number of galactic sources of the radiation but did not reveal any radiation from the atmosphere.

These observations mark the beginning of astronomical studies in ultraviolet radiation. The fact that so much has already been obtained in a single rocket flight lasting only a few minutes is an adequate testament of the great scope afforded for the development of ultraviolet and X-ray astronomy by the prospect of systematic observations from satellites. The possibilities are certainly very great indeed, particularly when it is realized that the wavelength range of observation may also be extended to the infrared and microwave regions.

VOL
4
196

9. Cosmic and Auroral Radiations

General Considerations

The study of cosmic radiation has proved remarkably fruitful for the most fundamental aspects of physics. Thus it was from such studies that the positron, the μ-, π- and K-mesons and the hyperons were discovered and many of their properties determined. There is, however, another equally important side which is rapidly developing. The origin of cosmic radiation in itself poses a difficult and challenging problem and any evidence which may be obtained by observation which might help to throw light on this is of obvious value. Among such observations are those which establish the existence of variations in primary cosmic ray intensity which may be correlated with sidereal or solar time, with solar activity and so on (see Elliot, 1952). In recent years such correlations with solar flares and with magnetic storms have been established. Intensive study of these relationships is required to clarify the nature of the link between the related phenomena. It seems clear already that under certain circumstances part of the radiation is emitted by the Sun. The energy distribution and composition of the irregular solar contribution is yet hardly known and this makes it difficult to consider the possible solar sources.

A further important question concerns the composition of the primary cosmic radiation, particularly as regards the so-called heavy component which consists of very energetic bare nuclei of elements such as iron. The relation between the nuclear composition of cosmic rays and that of the Universe as a whole has an important bearing on the origin of the rays. Furthermore, it is possible to place limits on the path length traversed by the rays so that the chance of breakup of the heavier nuclei in fragmentation collisions with interstellar matter will not be too great.

Because the cosmic radiation is composed of charged particles it is affected by magnetic fields and this may be utilized to probe the fields of the Earth and of interplanetary space. It is only now that data on the cosmic ray intensity at points widely distributed over the Earth's surface is becoming sufficiently extensive to determine whether the magnetic field which acts on cosmic radiation just outside the Earth differs appreciably from that which would be expected from the analysis of ground level magnetic observations. This question is still unresolved but remains of great importance.

Evidence about the magnetic fields in interplanetary space, during disturbed solar conditions, may be derived from a study of the cosmic ray variations associated with the solar phenomena. These are of two kinds, sudden increases accompanying a solar flare and decreases which normally follow the onset of a magnetic storm. The flare increases, when measured in terms of the intensity of the neutron component at sea level, may be very great indeed, up to 50 times the normal intensity or more. Most of this probably consists of soft radiation from the Sun. The time development of the increase and subsequent decay to normal is of great interest and indicates that, for some time after the sharp initial rise to a maximum, the rate of decay is so gradual that radiation must be diffusing back to the Earth from the surrounding interplanetary space. This is probably due to scattering from regions in which the magnetic field is substantially higher than normal. From detailed study of these effects it is therefore possible that our understanding of the nature and origin of these sporadic and irregularly distributed interplanetary fields will be clarified.

We are even less clear about the processes which give rise to the cosmic ray decreases. Here again it is likely that in some way the effect is due to modifications of the magnetic field in some region of interplanetary space.

The cosmic ray energy spectrum extends downwards from extremely high energies to about 1000 MeV. Below this there appears to be little radiation of cosmic origin. However, under disturbed conditions, the Sun emits corpuscular streams which are composed mainly of protons and electrons moving with the same velocity and in such concentration that the charge density in any macroscopic volume within the stream is effectively zero—a neutral ionized stream. The energy of the protons in these streams is much less than that of the least energetic cosmic ray particles, being as low as 3×10^4 eV. Nevertheless the consequences of the interaction of a stream with the Earth's magnetic field and atmosphere are profound, leading to magnetic and ionospheric storms and auroral displays. The detailed relations between these phenomena are very complicated and have not yet been worked out in detail. We shall refer again to these in §10. Meanwhile it is convenient to draw attention here to one aspect which was discovered by VAN ALLEN and his collaborators using rocket-borne instruments (MEREDITH, GOTTLIEB and VAN ALLEN, 1955). In some way which is far from clear the interaction between a solar corpuscular stream and the Earth's magnetic field appears to lead to bombardment of the Earth's atmosphere in the auroral zones by protons and by electrons with an energy of several 100 keV. The electrons, in collisions with atmospheric atoms and molecules, emit soft X-radiation as "bremsstrahlung". These soft X-rays have been observed in VAN ALLEN's experiments at altitudes of about 100 km in the auroral zones and their presence is not only a sign of auroral activity but also is an effective method of delineating the zone. The remarkable observations made from the Explorer satellites which are described in §13, p. 104, were essentially a continuation of these studies. They led to the discovery of the radiation belts around the Earth which are certainly connected with auroral phenomena.

It is clear that any further methods of studying the nature, composition and terrestrial interaction of solar corpuscular streams will be very useful and important. In times of great solar activity this frequent and copius emission from the Sun may well provide the irregular variations of magnetic field which produce the scattering effects on solar cosmic rays.

We shall now discuss briefly how the availability of artificial satellites as instrument containers can assist in unravelling some of the complexities we have outlined above. Although the precise magnetic field measurement from a satellite will be of great importance in this regard, we shall defer consideration of such measurements till the next Section and consider here only observations which may be made of radiation intensity and composition.

Possible Satellite Observations

By and large, the chief advantage of a satellite in cosmic ray research is that it makes possible the direct recording of primary cosmic radiation free from secondary rays due to atmospheric interaction. This is not quite true because of the "albedo" effect. Some of the secondary rays are emitted in the backward direction (i.e. radially outwards) and then bent back by the Earth's magnetic field when outside the atmosphere. Such rays appear among the primaries. With a satellite in an elliptic orbit

the height variation of intensity of the radiation could be studied and from this the albedo contribution estimated, as it alone would show a marked height dependence.

The advantage of observing the primary rays directly is only immediately apparent for those constituents of the radiation which interact with atmospheric nuclei at altitudes well above the limit of balloon flight (30 km)—mainly the heavy components. Otherwise balloon-borne instruments would appear to offer great advantages and have been used extensively in the past. However, even for the primary proton and alpha-particle constituents a satellite offers the prospect of continuous world-wide observations over considerable periods of time. This is a very important prospect. Returning to the heavy components, it is true that rockets are capable of carrying instruments up to the required altitudes for the study of these primary components but only for a time which is far too short to obtain useful data about relatively weak radiation.

The simplest observations which may be carried out by instruments within a small satellite are of total cosmic ray intensity. This may be done in principle by incorporating within the satellite one or more Geiger counters connected to an appropriate telemetry system. Suitable counters can be made which are very small, can withstand high accelerations and have a low power device. The presence of soft radiations, which can nevertheless actuate unshielded counters and which may be much more intense than the cosmic radiation (see §13), makes it necessary to include discriminating devices of some kind in order that cosmic radiation alone should be recorded. These devices may simply involve shielding or more elaborate methods using scintillation counters and pulse height analysis.

From observations made with these counters, together with reliable orbit data, it is possible to obtain the variation of total primary cosmic ray intensity with geomagnetic latitude (see NEHER, 1952). This in turn provides a means of determining the energy distribution of the cosmic rays. Thus the least possible momentum a particle of charge e must have in order to enter the atmosphere vertically at geomagnetic latitude ψ is given by

$$p = 14{\cdot}9 \cos^4\psi \; Gev/c.$$

As the latitude increases, more and more energetic particles are excluded so that the total intensity falls. This may be put on a quantitative basis to derive the energy distribution of the primaries. There is evidence that this varies from time to time particularly during and just after solar flares, but possibly at other times also. Continuous observations will therefore be of very great value.

Combination of such data with observations at different longitudes and directions of incidence will lead to information about the magnetic field felt by the primary rays.

VAN ALLEN (1956) proposed an experiment in which the intensity ratio of the Li, Be and B nuclei in the primary rays to that of nuclei with atomic number $Z \geqslant 7$ is to be studied (see p. 103). The particles are to be detected by the Čerenkov radiation which they produce in a block of Lucite. Discrimination is based on the length of the pulse produced by the Čerenkov radiation in a photomultiplier. Even including battery power it is considered that this apparatus need only weigh about 8 lb and in a week should detect about 1000 incoming nuclei with $Z \geqslant 6$. As it is only the intensity ratio which is being observed it is unnecessary to have accurate orbit or aspect data or continuous telemetry transmission.

Apart from the main question of the relative abundance of different components in the primary cosmic rays, VAN ALLEN's apparatus would also provide a very good means for answering another question. It seems quite certain that, during solar flares, the Sun may make a considerable contribution to the cosmic radiation, probably in the low energy range. What is the composition of this solar contribution? Although in time this could be answered by balloon studies it requires slow and tedious work, as the balloon carrying the necessary photographic emulsions or other detecting and discriminating equipment must be launched very soon after a flare on the Sun has been observed. This is much easier said than done and even when achieved gives less direct results than that of observation from a satellite.

There seems little doubt that the availability of satellite instrument carriers opens out exciting possibilities for cosmic ray research which are likely to be exploited for years to come.

The soft radiations, of the type observed by VAN ALLEN and his associates, with unshielded counter equipment flown in rockets above the auroral zones, can be investigated in a similar way from satellites. Here again a great deal more information may be obtained by using discriminating detectors so that the nature and energy distribution of the radiation may be studied. The observations made by VAN ALLEN with counters carried in Explorer satellites (see p. 104), which have revealed the existence of belts of intense radiation surrounding the Earth, are a first example of the discoveries in this field which may be made using satellite borne equipment, and have been followed by several other investigations which are described in §13, pp. 104–6.

10. MAGNETIC MEASUREMENTS

General Considerations—The Outstanding Problems

We now consider what interest attaches to precise measurements of magnetic field at altitudes well above the Earth's surface.

An immense amount of material relating to the magnetic field at the Earth's surface and its variations now exists (CHAPMAN and BARTELS, 1940). A great deal of time has been devoted to its analysis so that we are able to distinguish the main field and its secular variations from the quiet and disturbed variations. The quiet variations have been further separated into a predominantly solar and a weaker lunar component. Several regularities have even been discussed within the so-called disturbed variations and the general march of events in a magnetic storm is well established. Nevertheless, many difficult problems remain to be answered.

It is possible from a thorough survey of the Earth's main field at the surface to derive the field at any points above the surface, assuming that no contribution comes from local sources in the atmosphere. This may be done by carrying out a spherical harmonic analysis of the surface field based on available data. With the available coverage of magnetic observations, bearing in mind that there are large areas of sea over which few data on the surface field are available, it is not possible to represent the actual surface field to within 150 γ ($1{\cdot}5 \times 10^{-3}$ gauss). Errors of this magnitude are therefore likely to arise in interpolation, by means of the analysis, to higher altitudes. There is obviously room for much improvement here, apart from any question of measuring magnetic variations. With a satellite in a distant and hence

long-lived orbit it would even be possible to obtain more information about the secular variations.

Turning now to the quiet magnetic variations we can assume that they arise from dynamo currents produced in the ionosphere due to the solar and lunar tides causing drifts of ionized matter across the Earth's magnetic field. We are still uncertain about the altitudes at which these dynamo currents circulate. Moreover, the solar and lunar variations differ not only in period and amplitude but in other ways which suggest that they arise from current systems at different altitudes. Useful information on these matters could be obtained by measurement in a satellite which could be correlated with ground measurement at the same geomagnetic coordinates. Thus if the source of the field giving rise to a particular quiet variation is between the surface and the altitude of the satellite the sense of the variation of horizontal components will be opposite in the satellite-borne magnetometer to that at ground. In fact a more complicated analysis would be necessary as a satellite will not approach within 150 km except in the last stages of its life and most of the dynamo current flow will probably be below this altitude.

Even a rough localization in altitude would solve one of the major problems associated with the disturbance variations. A magnetic storm is normally characterized by a sudden commencement (SC) followed by an increase of horizontal component H which lasts for a few hours. This is succeeded by the main phase which consists of a sharp decrease in H followed by a gradual recovery to normal which may take several days. In practice, apart from irregular short-time fluctuations, there are superposed on the main phase, disturbances which have solar periods. These are normally referred to as the S_D variations to distinguish them from the quiet S_q and L_q. The main phase is denoted as the D_{st} field. The problem is to understand the origin of these disturbance fields.

There is a definite correlation between the onset of magnetic storms and of auroral displays. It has also been established that both are correlated with disturbed conditions in the Sun but with a time delay of about one day. These results are interpreted as indicating that the emission of a neutral ionized stream from the Sun is responsible both for the magnetic storms and the auroral phenomena. The time delay represents the time taken for the stream to reach the neighbourhood of the Earth. Unfortunately, the analysis of the interaction of the stream with the Earth's magnetic field is very complicated and it is impossible to trace the sequence of events by purely theoretical arguments. One of the most widely accepted theories, that of CHAPMAN and FERRARO (1940) suggests the following interpretation of a typical magnetic storm sequence.

As a corpuscular stream penetrates the Earth's magnetic field currents are induced in the front of the stream which are such as to neutralize the magnetic field within the stream. The first phase of the storm is then interpreted as arising from the field of the induced currents which will lead to an increase of horizontal component on the Earth's surface. The induced currents will be in such a sense that the forces on them due to the increasing magnetic field will tend to stop the further penetration of the stream into the field. In fact a part of the stream will be stopped if it enters a region where the magnetic field H is such that the magnetic pressure $H^2/8\pi$ is equal to the kinetic pressure $\frac{1}{2}nmV^2$ of the stream, m being the mass and V the velocity of the stream particles and n their concentration. Because of this the Earth will dig out a hollow in the front of the advancing stream. Up to this stage the argument is quite

definite but the subsequent stages are largely hypothetical. It is assumed (MARTYN, 1951) that as the stream passes beyond the Earth, a ring current is set up in the equatorial plane circulating around the Earth at several Earth radii distance. Owing to its high electromagnetic inertia the ring current will persist for several days. As it will give rise to a decrease in horizontal component of magnetic field at the Earth it is natural to associate it with the main phase of the storm.

There remain the S_D variations. These are considerably more difficult still to understand as they depend on the detailed nature of the interaction between the stream and the Earth. It is likely that they owe their origin to the production of current systems within the atmosphere due to bombardment by protons and electrons, primarily in the auroral zones. The complex nature of the phenomena is indicated by the fact that these charged particles, which in some way arise from the incident solar stream, must, in order to penetrate to the zones at which aurorae occur, be considerably more energetic than the stream particles themselves (the velocity of these latter particles is fixed by the observed time delay between the onset of solar disturbances and of the associated terrestrial phenomena).

It is quite clear from the above very incomplete sketch of the present state of our knowledge that it is of the first importance to locate the current systems responsible for the various effects. The ring current is sufficiently hypothetical to require much more definite observation and justification. Precise measurement from a satellite following an elliptic orbit would even check whether the D_{st} field largely arises from current systems located outside or inside the orbit and this would be a big step forward. There is even some doubt as to whether the first phase of a storm is really due to induced currents in the incident stream. An atmospheric origin could be detected quite readily from continuous satellite observations. Finally, any information which would bear on the origin of the S_D variations would be most welcome.

The discovery of the outer radiation belt surrounding the Earth, as described in §13, has already led to considerable further speculation about the origin of auroral displays. It seems certain that further systematic observations, both of magnetic fields and of the intensity and energy distribution of the particles in the belt, will throw fresh light on the whole subject.

The Technique and Problems of Magnetic Measurements from Satellites

When one considers what is required to make the desired magnetic measurements from a satellite it soon becomes apparent that many difficulties have to be overcome.

It will first be necessary to work with a satellite composed of strictly non-magnetic material. This does not offer insuperable difficulties as the satellite itself may be made of a suitable plastic.

Great emphasis must be laid on accurate tracking as the main field must always be subtracted off. HEPPNER (1956) has considered this question in detail. To take an example, for a satellite at 200 miles altitude an error of 5 miles in altitude determination would give rise to an error in the estimation of the main field as large as 100 γ which is as great as that to be expected in the D_{st} phase of a moderate magnetic storm. To reduce the error to 20 per cent of this value it would be necessary to know the altitude to within 1 mile. Correspondingly greater accuracy in altitude would be required to measure smaller disturbance fields. Thus, for a 20 γ disturbance at 30° magnetic altitude, the altitude would have to be known to within 0·2 miles in order

that the error in estimation of the main field should not be greater than 20 per cent of the disturbance field.

It is of course necessary in view of the present inadequate knowledge of the main field at satellite altitudes that this be investigated systematically with satellite-borne magnetometers of high precision, associated with very precise orbit determinations.

The kind of investigation which may be carried out depends considerably on the inclination of the orbit of the available satellites. Its scope depends also on the number and disposition of telemetry receiving stations and surface magnetic observatories as simultaneous measurements of field at two altitudes is one of the most powerful ways of interpreting the results.

A most important point is the choice of a suitable magnetometer. This must be capable of providing field data correct to a few gamma. It is very difficult to design an instrument which is capable of measuring field components from a satellite to this order of accuracy but the free proton precession magnetometer offers a suitable instrument for measurement of total field to well within the required accuracy.

The Proton-Free Precession Magnetometer

This instrument takes advantage of the fact, that in a magnetic field of strength H, the direction of the dipole moment of a free proton precesses about the field direction with a frequency

$$\nu_p = \gamma_p H/2\pi, \qquad \ldots . (10.1)$$

where γ_p, the gyromagnetic ratio for the proton, has been measured as $2{\cdot}6752 \pm 0{\cdot}002 \times 10^4\,\mathrm{sec}^{-1}\,\mathrm{gauss}^{-1}$. The precessing moments will induce an EMF in a suitably oriented pickup coil and the frequency of this EMF may be measured to give γ_p and hence H.

In the free precession instrument the protons in a suitable hydrogen-containing material are oriented along a strong polarizing magnetic field M, which is much stronger than, and nearly perpendicular to, the field H to be measured. The polarizing field is suddenly switched off, leaving the protons to precess coherently about the direction of the field H.

For the Earth's field the frequency ν_p is about 2 kc/sec. It is well known that frequency measurement can be performed more accurately than other physical measurements and there is no difficulty then in obtaining H from (10.1) to an absolute accuracy of 1 γ while making as many as 15 measurements per min. This is independent of field orientation. Furthermore, it is not necessary to include the apparatus for frequency measurement within the satellite. The signal from the pickup coil can be telemetred to ground and analysed for frequency by ground level instruments.

Concluding Remarks

Although there is no doubt that close attention must be paid to high precision of satellite location as well as of magnetic field measurement and that in any case elaborate analysis of new data will be required, there are great prospects of obtaining information of vital importance for geomagnetism by using satellite-borne magnetometers. The value of these data will be greatly enhanced if simultaneous observations are possible over a wide network of ground-based stations.

11. The Ionosphere

Introductory Remarks

A great variety of important investigations of the ionosphere may be carried out using artificial satellites. They may be divided into two distinct categories. The first consists of propagation studies which simply require a radio-beacon in the satellite whose signals may be recorded at the ground. Using suitable receiving equipment it is then possible to obtain information about the electron distribution in the ionosphere without requiring knowledge of the instruments within the satellite. In this way countries which have not themselves launched satellites may make most important and largely self-contained observations. It is even possible, if radar tracking can be carried out, to obtain, from the characteristics of the reflected signal, information about the electron content of the intervening medium.

The second category involves observations of ionospheric properties by means of instruments contained within the satellite itself. This includes, for example, the determination of the positive ion composition, the magnitude of any electric fields which may be present, etc. In these cases, the technique is similar in principle to that used in the experiments discussed in §7, p. 67.

Propagation Studies

We consider first the type of information which can be obtained from reception of the radio signals transmitted from the beacon or beacons in the satellite. It turns out that this mainly yields the electron concentration integrated along the propagation path from satellite to receiver. Thus if n_e is the electron concentration one obtains

$$\int_0^s n_e \mathrm{d}s \qquad \ldots\ldots (11.1)$$

where $\mathrm{d}s$ is an element of length on the propagation path which is measured from the satellite, s being the distance along the path to the receiver. If the frequency is well above the critical frequency for the ionospheric regions through which the ray passes, the analysis of the data is rendered much simpler and one obtains essentially

$$\int_0^h n_e \mathrm{d}h \qquad \ldots\ldots (11.2)$$

where h is the height of the satellite above the receiver, and the integration is along the vertical.

Under certain conditions it is possible to determine the actual electron concentration near the satellite but even if this proves difficult the fact that a satellite pursues an elliptic orbit means that the integral in (11.2) may be obtained for different values of h, so that more specific results about the electron distribution in height may be obtained.

Although a satellite will only pass through the E-region during the last stages of its life it may circulate within the F-region for long periods so that the prospect is opened of great extension of our knowledge of this region from satellite propagation studies. The F-region is known to be one in which the behaviour is greatly complicated by electrodynamic effects which give rise to vertical drifts of ionization.

Because of this, the behaviour of the level of maximum electron concentration is very complicated as regards variation of height and altitude. In principle, the electrodynamic effect may be averaged out by considering the variation of the total electron content of the region at a particular latitude and longitude. Such variations will be determined solely by processes of electron production and loss. It is not possible, however, to determine the total electron content from ground level ionospheric probing methods. They only serve to give the concentration up to the layer maximum. Nevertheless Ratcliffe and his collaborators (1956) have shown that, by working with the total concentration up to the maximum, the complications due to the electrodynamic effect are much reduced. The availability of radio-sources within the F-layer, at altitudes extending beyond the layer maximum out to great distances, now opens the prospect of a complete removal of electrodynamic complications, by providing data on the total electron content of the region over a wide range of time and position.

Apart from this special interest from the point of view of checking the current theory of the processes of electron production and loss which are important for the region, the new information which will become available about the electron concentration above the layer maximum will be of much interest in itself, and may reveal greater complexity in the height distribution than is at present suspected.

The moving source of radio waves above or within the ionosphere may also be used to study ionospheric irregularities, including their distribution in height.

A specially interesting possibility arises when radar reflection from the satellite is observable. From the polarization of the received signal (the Faraday effect described below) it is possible to derive the integrated electron concentration between the transmitter and target. In the same way by reflecting from the Moon this concentration integrated between the transmitter and the Moon can be derived. The difference gives the integrated concentration between the satellite and the Moon. This is a technique capable of yielding most important information about interplanetary space.

The usefulness of the propagation studies using the radio beacon depends on the modifying influence of the ionosphere on the propagation. In order that this may be derived with the necessary accuracy it is necessary to know the true position of the satellite, at any time when reception is taking place, to a high precision from optical observations.

We now consider in more detail the methods which may be used to derive ionospheric information from the nature of the received signals. Three effects of the ionosphere may be used: refraction, modification of the Doppler effect of the signals and the rotation of the plane of polarization (Faraday effect).

We shall discuss in detail only relatively simple cases which illustrate the principles involved. Thus it will be assumed that the radio frequencies concerned are great compared with the critical frequency for the ionospheric regions beneath the satellite. The ionosphere will be regarded as horizontally stratified and the Earth's curvature will be neglected. We must first summarize some of the relevant features of the magneto-ionic theory of ionospheric propagation.

Formulae of the Magneto-Ionic Theory

Due to the presence of the Earth's magnetic field, a region in which the electron concentration is n_e is a doubly refracting medium for radio waves. For a ray of

frequency f propagating in a direction making an angle ϕ with that of the Earth's magnetic field, the two refractive indices μ are given by

$$\mu^2-1 = \left[-\frac{f^2}{f_0^2}\left(1+\frac{f_h^2\cos^2\phi}{2(f_0^2-f^2)}\right)\pm\frac{ff_h}{f_0^2}\left(\frac{f^2f_h^2\sin^2\phi}{4(f_0^2-f^2)^2}+\cos^2\phi\right)^{\frac{1}{2}}\right]^{-1} \quad \ldots\ldots (11.3)$$

The frequencies f_0 and f_h are given by

$$f_0^2 = \frac{n_e e^2}{\pi m}, \quad f_h = \frac{eH}{2\pi mc}, \qquad \ldots\ldots (11.4)$$

m being the mass and e the charge of an electron and H the strength of the Earth's magnetic field. Of the alternative signs the upper one corresponds to the ordinary, the lower to the extraordinary ray respectively. This formula ignores the effect of collision damping, a valid approximation in dealing with the F-layer.

The Larmor frequency f_h is less than f_0 in the F-layer so that, if the frequency f is much greater than f_0, as we are assuming,

$$\mu^2-1 \simeq -\frac{f_0^2}{f^2}\left[1\pm\frac{f_h}{f}\cos\phi\right]^{-1}. \qquad \ldots\ldots (11.5)$$

The refractive indices μ_0, μ_e of the ordinary and extraordinary rays are therefore both nearly equal to

$$\mu \simeq 1-\frac{1}{2}\frac{f_0^2}{f^2} \qquad \ldots\ldots (11.6)$$

while the difference

$$\mu_0-\mu_e = \frac{f_0^2 f_h}{f^3}\cos\phi. \qquad \ldots\ldots (11.7)$$

Let h_z be the component of the magnetic vector in the wave which lies in the plane containing the wave normal and the direction of the Earth's magnetic field, h_y the perpendicular component so that the ratio h_z/h_y specifies the state of polarization of the wave. For a plane polarized incident wave entering the medium we find that within the medium

$$h_z/h_y = -[g\pm\sqrt{1+g^2}], \qquad \ldots\ldots (11.8)$$

where

$$g = \frac{ff_h}{2(f^2-f_0^2)}\tan^2\phi, \qquad \ldots\ldots (11.9)$$

the alternative signs corresponding to the two rays. This shows that the two rays are elliptically polarized with major axes at right angles and opposite senses of circulation. If f is much greater than f_0 and f_h the ellipses are effectively circles. At any point the resultant of the two circularly polarized waves is a linearly polarized one but with a plane of polarization depending on the distance traversed through the medium.

Thus if λ_0 and λ_e are the wavelengths $c/\mu_0 f$, $c/\mu_e f$ of the two waves, the respective magnetic vectors will be rotated through angles $2\pi\delta s/\lambda_0$, $2\pi\delta s/\lambda_e$ in opposite senses

in traversing a distance δs through the medium. This will produce a rotation of the plane of polarization by an amount

$$\begin{aligned}\delta\theta &= \pi\delta s\left(\frac{1}{\lambda_0}-\frac{1}{\lambda_e}\right)\\ &= \pi\delta s f(\mu_0-\mu_e)/c,\\ &\simeq \frac{\pi f_0^2 f_h}{cf^2}\cos\phi\,\delta s, \quad \text{from (11.7).} \qquad \ldots.(11.10)\end{aligned}$$

Hence, if the electron concentration varies with s, the rotation of the plane of polarization due to passage from s_1 to s_2 through the medium along the ray path is given by

$$\theta = (\pi f_h/cf^2)\int_{s_1}^{s_2} f_0^2\cos\phi\,\mathrm{d}s,$$

$$= \frac{e^3H}{2\pi m^2c^2f^2}\int_{s_1}^{s_2} n_e\cos\phi\,\mathrm{d}s. \qquad \ldots.(11.11)$$

If n_e is horizontally stratified it is convenient to rewrite this in terms of an integration over height h so that

$$\theta = \frac{e^3H}{2\pi m^2c^2f^2}\int_{h_1}^{h_2} n_e(h)\cos\phi(h)\frac{\mathrm{d}s}{\mathrm{d}h}\mathrm{d}h. \qquad \ldots.(11.12)$$

This is the Faraday rotation which, if it can be measured, enables one to obtain information about the total electron content along the ray path, particularly when $\phi(h)$ and $\mathrm{d}s/\mathrm{d}h$ do not vary appreciably over the range of h concerned.

Application to Satellite Observations

We now consider the actual application to satellite observations (WEEKES, 1958). In principle ionospheric refraction is the simplest phenomenon. This leads to a difference in the apparent bearing of the satellite, as determined from the direction of the ionizing radio signal, and the true bearing.

It will be found that the ionospheric contributions to refraction and the Doppler effect are quite closely related. We neglect the curvature of the Earth and regard the ionosphere as horizontally stratified. Consider the path of a ray from the satellite at S to a point O on the surface of the Earth where it arrives at an angle of incidence i (Fig. 16). If h_s is the altitude of the satellite and x is the horizontal distance ON to the vertical SN from the satellite we have

$$x = h_s\tan i_0, \qquad \ldots.(11.13)$$

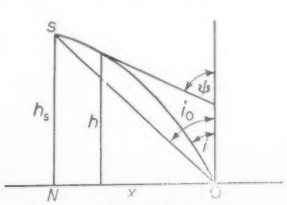


FIG. 16.

where i_0 is the "true" angle of incidence of the straight line SO. If ψ is the angle the tangent to the ray makes with the vertical when at an altitude h then

$$\mu \sin\psi = \sin i, \qquad \ldots . (11.14)$$

where μ is the refractive index at the altitude h. Since

$$x = \int_0^{h_s} \tan\psi \mathrm{d}h, \qquad \ldots . (11.15)$$

we have

$$h_s \tan i_0 = \int_0^{h_s} \frac{\mathrm{d}h}{(\mu^2 - \sin^2 i)^{\frac{1}{2}}},$$

or

$$h_s \tan i_0 = \tan i \int_0^{h_s} \frac{\mathrm{d}h}{\{(\mu^2-1)\sec^2 i + 1\}^{\frac{1}{2}}} \qquad \ldots . (11.16)$$

At the frequencies in which we are interested $\mu^2 \ll 1$ so that, except for angles of incidence very close to $\pi/2$,

$$h_s \tan i_0 = \tan i \int_0^{h_s} \{1 - \tfrac{1}{2}(\mu^2 - 1)\sec^2 i\}\mathrm{d}h,$$

or

$$\tan i_0 = \tan i \left[1 - \frac{1}{2}\frac{\sec^2 i}{h_s}\int_0^{h_s} (\mu^2 - 1)\mathrm{d}h\right]. \qquad \ldots . (11.17)$$

Thus if $i = i_0 + \delta i_0$ where δi_0 is small we have, if i_0 is not too close to $\pi/2$,

$$\delta i_0 \simeq \frac{\tan i_0}{2h_s}\int_0^{h_s} (\mu^2 - 1)\mathrm{d}h$$

$$= \frac{e^2 \tan i_0}{2\pi m h_s f^2}\int_0^{h_s} n_e \mathrm{d}h, \quad \text{from (11.5)} \qquad \ldots . (11.18)$$

and depends on the integrated electron concentration up to the satellite (11.18).

To estimate the magnitude to be expected for δi_0 we suppose the satellite to be at an altitude of 500 km and take

$$n_e = 5{\cdot}5 \times 10^5/\mathrm{cm}^3 \ , \quad h > 300 \text{ km},$$
$$= 0 \qquad\qquad\quad , \quad h < 300 \text{ km}.$$

(11.18) then gives, for radiation of frequency 20 Mc/sec and $i_0 = 45°$,

$$\delta i_0 \simeq 0{\cdot}02 \text{ radians}.$$

The effect of the ionosphere on the Doppler shift is a little more complicated. A similar situation arises in the method for tracking high altitude rockets based on the Doppler shift of signals from a radio beacon carried in the rocket. If the rocket passes through the ionosphere errors are introduced. A knowledge of these errors (as for example from comparison of the Doppler-determined trajectory with the true one obtained in some alternative way) may be used to obtain information about the ionosphere.

The Doppler effect arises from a change in the number of waves along the propagation path between the source and the observer. If the ray path is effectively straight, this number is given by

$$\int_0^{s_1} \frac{\mathrm{d}s}{\lambda(s)} \qquad \ldots\ldots(11.19)$$

where s is measured along the ray path of length s_1 from the observer to the satellite. $\lambda(s)$, the wavelength at a distance s from the observer, is given by $c/f\mu(s)$ where f is the natural frequency, c the velocity in free space and $\mu(s)$ the refractive index at a distance s.

The Doppler shift in frequency is then given by

$$\Delta f = \frac{f}{c}\frac{\mathrm{d}}{\mathrm{d}t}\int_0^{s_1} \mu(s)\mathrm{d}s. \qquad \ldots\ldots(11.20)$$

We may now evaluate (11.20) by the same procedure as for refraction

$$\begin{aligned}\int_0^{s_1} \mu(s)\mathrm{d}s &= \int_0^{h_s} \mu(h)\sec\psi\mathrm{d}h \\ &= \int_0^{h_s} \frac{\mathrm{d}h}{(\mu^2-\sin^2 i)^{\frac{1}{2}}} \\ &\simeq h_s\sec i - \frac{e^2\sec^3 i}{2\pi m f^2}\int_0^{h_s} n_e\mathrm{d}h. \qquad \ldots\ldots(11.21)\end{aligned}$$

Since $i \simeq i_0$ we then have, using (11.18),

$$\int_0^{s} \mu(s)\mathrm{d}s = h_s\sec i_0 - \frac{e^2\sec i_0}{2\pi m f^2}\int_0^{h_s} n_e\mathrm{d}h, \qquad \ldots\ldots(11.22)$$

so that the Doppler shift in frequency Δf is given by

$$\Delta f = \Delta f_0 - \frac{e^2}{2\pi m c f}\left\{\dot{h}_s\sec i_0 n_e(h_s) + \frac{\mathrm{d}}{\mathrm{d}t}(\sec i_0)\int_0^{h_s} n_e\mathrm{d}h\right\}. \qquad \ldots\ldots(11.23)$$

Δf_0 is the shift which would occur in the absence of ionospheric effects. The second

term depends only on the electron concentration at the satellite while the third again involves the integrated concentration. Under the same conditions as those used in estimating the refraction effect the contribution of the third term is about 12 c/sec. The second term is much smaller except when i_0 is very close to $\pi/2$. It is possible, in principle, by a combination of Doppler and refraction measurements, together with accurate knowledge of true position, to determine n_e as well as

$$\int_0^{h_s} n_e \mathrm{d}h$$

but in practice this is difficult because of the high accuracy required. This involves measurement of the Doppler frequency to about 0·5 c/sec and the value of $\sin i$ to about 0·1 per cent or i to 2′ of arc. Otherwise, analysis of the Doppler effect and of refraction give the same information.

Turning now to the use of the Faraday effect we see that, with a linearly polarized receiving aerial, the received signals will show fading due to the change of the Faraday rotation with the satellite motion. From the fading rate the rate of change of θ in (11.12) may be obtained and knowing the position and velocity of the satellite throughout the time of observation it is possible to derive a weighted mean of the electron concentration up to the height of the satellite.

In practice there will usually be other sources of fading as well as that due to the Faraday effect. Separation of the different contributions is much facilitated if observations can be made at two different frequencies as the Faraday effect dependence on the square of the frequency may be a sufficient distinction (see §13, p.101).

If radar reflection from the satellite can be observed and the rotation of the plane of polarization measured (11.12) may be used directly to obtain information about the weighted mean of the electron concentration between source and satellite.

Applications of these methods to observations of the first two artificial satellites will be discussed in §13, p. 101.

Ionospheric Measurements in situ

We have already discussed (§7, p. 68) one device, the omegatron, which may be used in a satellite to obtain information about the relative concentrations of different ionized and neutral species in the ionosphere. The use of other types of mass spectrometer is also practicable. The possibility of carrying out a probe analysis of the ionospheric plasma has been raised by Hok, Sicinski and Spencer (1957) and by Gringauz and Zelikman (1957).

The latter authors point out that, if a Langmuir probe analysis is undertaken, the time required to scan the whole characteristic in such a way as to permit telemetering to ground may to be several seconds, during which the satellite will move several hundred metres through the ionosphere. This distance is large enough for ionospheric irregularities to appear so that the characteristic could be seriously distorted. They suggest instead that a double probe be used in the following way.

Let ϕ be the electrostatic potential of a small area of the satellite surface relative to the surrounding ionospheric plasma. The positive ion current per unit area which reaches this surface element will be given by

$$j^+ = en^+ v_s \cos \vartheta f(\phi), \qquad \ldots . (11.24)$$

where n^+ is the positive ion concentration, v_s the satellite speed and ϑ the angle the direction of motion which the satellite makes with the normal to the surface element. $f(\phi)$ is a correcting factor, which arises because of acceleration or deceleration of the ions by the electrostatic field across the sheath between the satellite surface and the plasma.

In obtaining this expression account is taken of the fact that the speed of the satellite is great compared with the mean thermal speed of the ions. Because of this there will be a space behind the satellite which will be devoid of positive ions and the expression (11.24) will not be valid if the small area concerned faces this region.

The correcting factor $f(\phi)$ will not differ much from unity if

$$e\phi \ll \tfrac{1}{2}mv_s^2$$

where m is the mass of a positive ion. For a nitrogen ion N_2^+ the right-hand side is about 9·5 eV for a typical satellite speed v_s. When the satellite is not illuminated it is probable that ϕ is a few volts negative at most and may be less than 1 V. Under these conditions $f(\phi)$ will not differ much from 1. During daytime the situation is much less clear (see §3, p. 39) and it would be of considerable interest to be able to measure ϕ.

The suggested equipment for measuring j^+ and ϕ is as follows. Two spherical collectors project symmetrically from diametrically opposite points of the satellite (assumed spherical) so that one will be within the ion-free region behind the moving satellite. The resistance of the connection from the collectors to the conducting surface envelope of the satellite is of the order of $10^6\Omega$. Each sphere is surrounded by concentric spherical grids insulated from the collectors but connected to the satellite envelope through a resistance of order $10^3\Omega$.

To measure j^+ the grids are maintained at the same potential as the satellite surface but the collectors are kept at a negative potential relative to the grids. This is sufficiently high to prevent any electrons from reaching the collector yet not high enough to cause any appreciable secondary emission from the collectors by positive ion impact.

Measurement of ϕ is to be carried out by observing the characteristic variation of ion current to the collector with potential difference between the spherical grids and the satellite surface. The voltage swing would be provided by means of a generator which delivers, at intervals of a few seconds, saw-tooth voltage impulses of both signs with amplitude 20 volts or so and a duration of about 0·2 sec. The generator would be connected between the grids and the satellite surface. As the grids become more and more positive with respect to the satellite surface a point will be reached at which all positive ions will be prevented from reaching a grid. This will occur when ϕ_g, the potential difference between the grid and the plasma is given by

$$\phi_g = \tfrac{1}{2}mv_s^2 \qquad \ldots . (11.25)$$

where m is the mass of a positive ion. Knowing the potential difference $\Delta\phi$ between the grid and the satellite envelope at this point it follows that $\phi = \phi_g - \Delta\phi$. The point on the ($j^+$, $\Delta\phi$) characteristic at which (11.25) is satisfied should be identifiable as indicated on Fig. 17. Increase of $\Delta\phi$ beyond the point should lead to no further reduction in positive ion current to the collector. In practice the identification may be obscured to some extent because the positive ions will not all be of the same mass.

Results obtained by the use of this equipment in Sputnik III are described in §13, p. 88.

Imyanitov (1957) has considered the problem of measuring electrostatic fields in the ionosphere using equipment carried in a satellite. This is a difficult problem in principle as the fields in the sheath at the satellite surface are likely to be very much

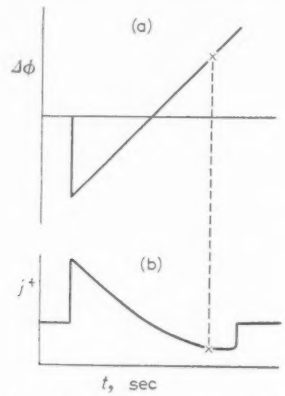


Fig. 17. Illustrating the proposed method for determining the potential of a satellite relative to its surroundings. The upper curve (a) gives the time variation of the potential difference $\Delta\phi$ between the receiver and the satellite surface during a pulse while the lower curve (b) gives the corresponding variation of positive ion current j^+ to the receiver. At the point marked X the potential of the receiver relative to the surroundings is given by the condition (11.25).

greater than those in the surrounding medium. It seems that investigations of this kind might well await the result of observations of satellite potentials by some such method as that outlined above.

The difficulty of this kind of investigation has already been made clear in earlier attempts to study the ionic constitution of the ionosphere using rocket-borne mass spectrographs. Very variable results have been obtained which almost certainly do not represent the true ionic composition. Part of the trouble is undoubtedly due to contamination by gases from the rocket motors but part is also likely to arise from the existence of a considerable potential difference between the rocket and the surroundings. The fact that the speed of a satellite is high compared with the mean thermal speeds of the atmospheric atoms, molecules and ions is, however, a simplifying factor.

12. Study of Micrometeorites

The solid matter entering the Earth's atmosphere from outside may be classified in terms of their interaction with the atmosphere and the methods which are available for observing them.

Meteors with mass greater than a few tens of kilograms are too large to be evaporated before reaching the ground. At the other extreme are those particles known as micrometeorites which are stopped within the atmosphere without undergoing any significant physical change. These are particles which have a large surface-to-mass ratio and hence are able to radiate away rapidly the heat energy which is generated in their interaction with the atmosphere. Micrometeorites have masses less than 10^{-8} g and a radius of a few microns or less.

Particles of intermediate mass are evaporated in the atmosphere. The evaporated atoms have sufficient energy, because they possess the meteor speed, to be excited

and ionized by impact with atmospheric atoms and molecules. The excitation leads to the emission of the characteristic light of a meteor trail so making it possible to observe the trail either visually or photographically. The magnitude range (on the usual stellar scale) which may be covered in this way extends down to about +10. Such observations are restricted because of the difficulty of maintaining a regular watch for meteors and the need for good visibility. The reflection of radio waves from the ionized trail which is left by a meteor greatly extends the scope of observation and is available for meteors of magnitude down to +12.

By considering the mechanism of meteor ionization and excitation it is possible to relate the visual magnitude to the mass, and, by assuming a mean density and spherical shape, the mean radius. For a meteor of magnitude +12 the mass is about 10^{-6} g and the radius about 50 microns.

As a result of the visual and radio data we now have quite good information about the flux of meteors with visual magnitude less than +12, but the position regarding micrometeorites is very unclear. One method of estimating their flux is to assume that the result which holds for meteors, that the total mass reaching the Earth is nearly the same in each magnitude range, remains valid for micrometeorites. This very uncertain extrapolation is not in accord with evidence derived by VAN DE HULST (1947) from the study of the F-corona nor is it consistent with the admittedly very rough estimates based on the accumulation of deep-sea sediments.

The geophysical importance of micrometeorites may be considerable, particularly when account is taken of the likelihood that they will be electrically charged. It is therefore most desirable that information be gained about their flux and their mass, energy and charge distribution. As it is by no means certain that they are distributed isotropically in the space around the Earth it is a great advantage to carry out observations from satellites so that the geographical distribution and any time variations of flux may be studied. Even in connection with the technique of satellite observation it is important to know more about micrometeorites. Any surface exposed for one purpose or another from a satellite will be subject to bombardment by this dust. This will modify the surface properties so that the time of effectiveness of an experiment which depends on the maintenance of surface properties depends on the magnitude of the micrometeorite flux.

Many of the earlier satellites will include among their instruments an erosion gauge designed to measure the rate of erosion of a surface from the change of electrical resistance of the surface.

This will give information of direct value to the designers of instruments for satellite investigations but it is difficult to infer from it anything very definite about the actual concentration of micrometeoritic dust. The first step in this direction is to use the technique already developed for use in high altitude rockets—that of the impact microphone. This is quite effective for counting micrometeoritic impacts. For example in the Aerobee rocket flight of 1955 September 14 a crystal microphone was mounted against a diaphragm of stainless steel 0·017 in. thick and 25 in^2 area. The signal due to an impact was amplified before telemetry transmission. The response of the system was a maximum at about 70 kc/sec frequency and the amplifier had a 10 kc/sec band pass.

Calibration of the instrument was carried out by dropping graded sand from a height of about 2 cm and recording the response. Assuming this to be determined largely by the momentum of the impinging particle the tests indicated that the equipment

possessed adequate sensitivity to detect dust particles, of mass down to 10^{-11} g, before they are slowed down by the atmosphere. As the pressure of the solar radiation removes particles with mass 10^{-13} g or less from the neighbourhood of the solar system it seems that a great deal of the unexplored mass range can be studied with similar equipment.

An alternative mode of detection of micrometeoritic impact has been developed. This makes use of the fact that, on impact, some radiation is emitted. Although this is very weak it may be recorded by a photomultiplier.

The determination of mass and energy distributions is much more difficult and is hampered by a lack of knowledge of the physical phenomena which occur on impact. More laboratory study in this direction is required.

13. The Satellites already Launched

Up till the beginning of 1959 eight satellites had been successfully launched or perhaps more accurately, fourteen objects had been placed in orbits. These are listed in historical order in Table 3 which also summarizes some basic properties both of the objects and of their orbits. (In July 1961 the number of objects was about fifty.)

Certain general differences between the Russian and American objects are clear. Thus, in general, the Russian objects circulate in orbits of quite high inclination for which the rate of rotation of the perigee (see p. 53) is very small, whereas the American revolve in orbits of much smaller inclination so that the rate of rotation of the perigee is relatively rapid. The difference in mass between the two sets of objects is also noticeable. A further distinction is in the choice of radio-beacon frequencies. It will soon be apparent that quite a variety of information has been forthcoming from this first batch of artificial satellites.

It is of interest to note the different arrangements made concerning the separation of the final stage rocket case from the satellite proper. This separation was carried out in the first and third Russian launching and also in that of the Vanguard test sphere so that in both cases two distinct objects, the satellite proper and the rocket case, were placed in orbits. With Sputniks I and III the nose-cone protecting the satellite on passage through the lower atmosphere was also carried to the launching altitude and separated slowly from the satellite sphere. In the Vanguard scheme the protecting cone is removed before the launching altitude is reached. No separation between rocket case and satellite proper was attempted for the remaining vehicles, Sputnik II, Explorers I, III and IV, and Atlas. While this has the advantage of making the objects much easier to track optically the effect of air drag is more complicated because of the possibility of aspect variation. This makes it more difficult to determine, from the observed change of period with time, such important quantities as the air density and perigee altitude.

At the time of writing a great deal of attention has been paid to the analysis of data obtained by radio, radar and optical observation of the orbits of the different satellites. New and interesting results have been obtained about the air density at 200 km altitude and above, about the figure of the Earth and about the electron concentration above the maximum of the F-layer. We shall discuss this work first and then proceed to consider some of the results obtained from measurements made by instruments carried in the satellites.

Table 3

Satellites launched before 1959 January 1.

Satellite	*Launching date*	*Weight (lb)*	*Shape and size*	*Inclination of orbit*	*Initial*					*Radio transmitter*		*Date of destruction*
					Perigee distance (km)	*Apogee distance (km)*	*Period of revolution (min)*	*Rate of regression of line of nodes (deg/day)*	*Rate of rotation of perigee (deg/day)*	*Frequency (Mc/sec)*	*Power (mW)*	
Sputnik I (satellite sphere)	10 Oct. 1957	184	sphere 23 in. dia.	65·1°	226	951	96·2	3·17	−0·40	20 40	1000	1958 Jan. 4
Sputnik I (rocket case)	10 Oct. 1957	?	?	65·1	226	951	96·2			—		1957 Dec. 1
Sputnik I (nose cone)	10 Oct. 1957	?	?	65·1	226	951	96·2			—		
Sputnik II	9 Nov. 1957	1120	?	65·29	226	1671	103·7	2·62	−0·42	20 40	1000	1958Apr.14
Explorer I	31 Jan. 1958	31	cylinder 80 in. long 6 in. dia.	33·2	354	2555	114·81	4·27	6·48	108·03 108·00	60 10	
Vanguard I (test sphere)	17 Mar. 1958	3	sphere 6½ in. dia.	34·3	650	3958	134·18	3·01	4·40	108·03 108·00	5 10	
Vanguard I (rocket case)	17 Mar. 1958	50		34·3	650	3958	138·5					
Explorer III	26 Mar. 1958	31	cylinder 80 in. long 6 in. dia.	33·3	187	2801	115·7	4·19	6·22	108·03 108·00	60 10	1958Jun.28
Sputnik III satellite	15 May 1958	2926	conical 148 in. long 68 in. dia.	65·18	226	1873	105·97	2·49	−0·40	20·005		
Sputnik III (rocket case)	15 May 1958	?		65·19	226	1873	105·9					1958 Dec. 3
Sputnik III (nose cone)	15 May 1958											
Sputnik III (side shields)	15 May 1958											
Explorer IV	26 July 1958	38	cylinder 80 in. long 6 in. dia.	50·3	263	2211	110·18	3·55	2·89	108·03 108·00	30 10	
Atlas	18 Dec. 1958	8700		32·3	185	1482	101·47	5·62	8·56	107·97 107·94		1959 Jan 21

Mr. Harold S. W. Massey

Observations of the First Satellites and their Analysis

There has been a gradual improvement in the scope and accuracy of tracking since the first Russian satellites were launched. Very few precise optical observations were made for Sputnik I or its rocket case as, at the time, the organization of precision optical tracking stations initiated in connection with the Vanguard programme was still in its early stages. The use of kinetheodolites, already available for other purposes, was soon introduced and yielded very useful information indeed for Sputnik II and Sputnik III. Baker-Nunn camera stations have gradually come into operation so that the supply of precision optical data is increasing rapidly.

Because of the inadequate facilities for optical tracking of the early satellites, special emphasis was placed on radio observations. Even here there was need for rapid improvisation as all facilities had been organized on the assumption that the frequency would be 108 Mc/sec as for Vanguard. Furthermore, it was known in Britain that the Vanguard satellites would not pass over the country and there was little advance information about the Russian satellites.

However, very shortly after the first Russian satellite was launched, observations of the Doppler frequency shifts exhibited at each passage over Britain were made by a number of stations as well as by many radio amateurs. Radio interferometers operating at the appropriate frequencies were employed at the Mullard Laboratory, Cambridge, and the Royal Aircraft Establishment to obtain apparent angular position data accurate to about 1′ of arc, while direction-finding equipment was in operation at other stations (the Admiralty Signals Establishment and the Radio Research Station). Measurements of signal strength were made and many fading records obtained.

The actual satellite sphere of Sputnik I which carried the radio transmitters proved to be a very difficult object to track optically as it was as faint as a 6th magnitude star and, of course, was moving very fast. On the other hand the rocket case of Sputnik I was quite a bright object, being comparable with a star of magnitude − 1. A sample photograph of a track of the rocket case passing over Edinburgh is shown in Fig. 18a.

The radio transmitters in the satellite sphere of Sputnik I ceased to operate after 2 weeks so that the further (optical) observations were mainly confined to the rocket case. Fortunately, the overlap between radio observation of the sphere and optical observation of the rocket case during the first two weeks made it possible to use more effectively the later data on the orbit of the rocket case.

Great assistance in tracking both the sphere and the rocket case was afforded by the use of radar methods. The 240 ft. radio telescope at the Jodrell Bank Station of the University of Manchester, operating on frequencies of 36 and 120 Mc/sec with a beam-width of 8°, was particularly effective and succeeded, not only in tracking the rocket case, but also observing the satellite sphere itself. Figure 18c reproduces a record of a radar reflection obtained from the rocket case. Measurements were made of the reflecting areas of both these objects, the values found being 10 m^2 for the rocket case and $\frac{1}{2}$ m^2 for the sphere. Radar observations of the rocket case were also made at a frequency of 3000 Mc/sec using the 45 ft. radio telescope at the Royal Radar Establishment, Malvern, with a beam width of $\frac{1}{2}$°.

The radio transmitter in Sputnik II only operated for 1 week but more accurate methods of optical tracking came into operation during its lifetime so that useful

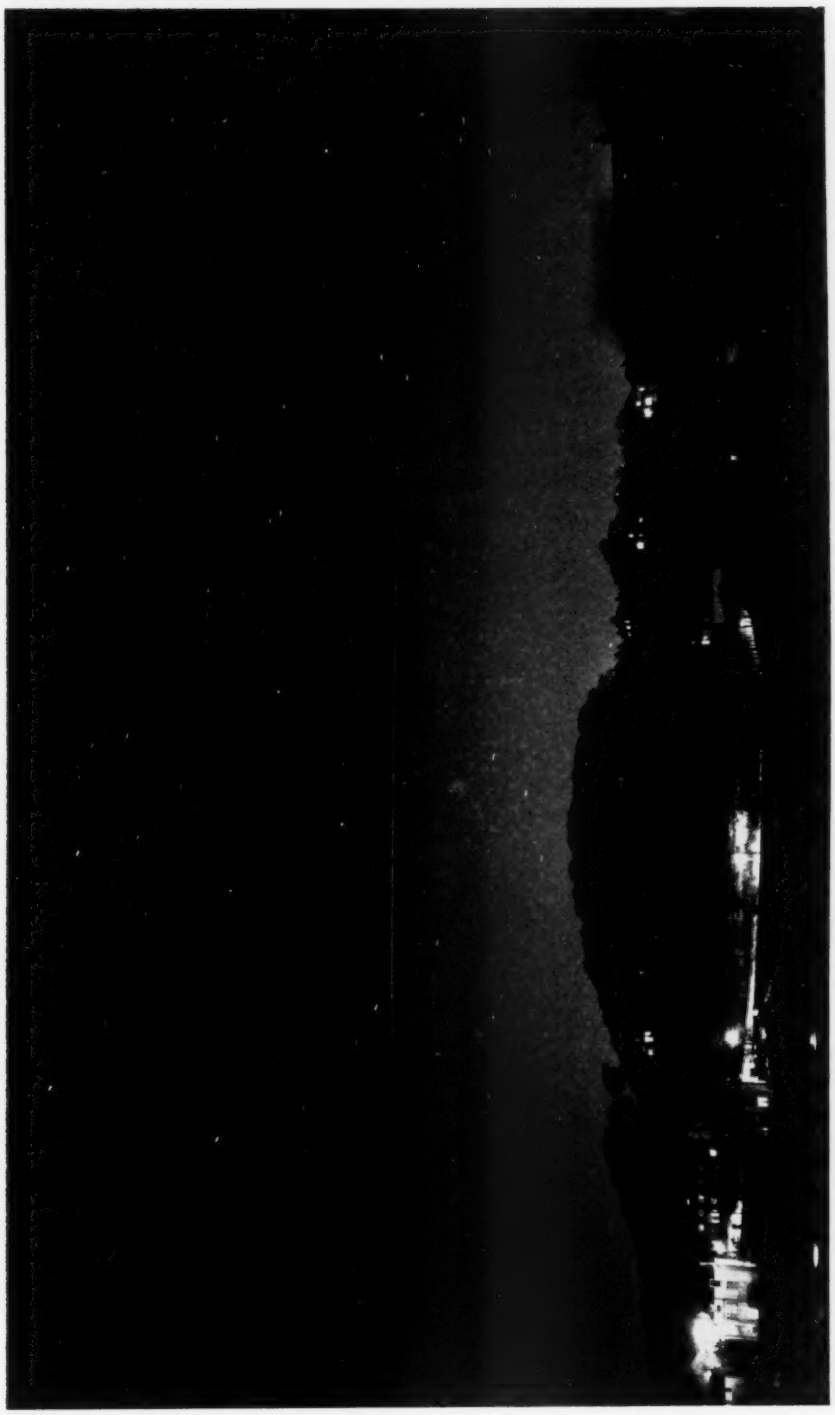


FIG. 18 (*a*)

Photograph of the trail of the rocket case of Sputnik I, looking north from the north shore of the River Forth at 18.20 hours local time on 1957 November 4. This photograph has been kindly supplied by MORRIS ALLEN of Dunfermline, Fife.

FIG. 18 (*b*)

Photograph of the track of Sputnik II, taken at the Royal Observatory, Edinburgh on 1958 January 17. The movement is from West to East (left to right). The timing cut-off at the two right-hand gaps in the trace correspond to times of 18 hr 24 min 38·0 ± 0·5 sec and 18 hr 24 min 48·0 ± 0·5 sec. This photograph (Ilford HP3, *f*/16, *f.l.* 8") has been kindly supplied by Dr. H. E. BUTTLER.

VOL
4
196

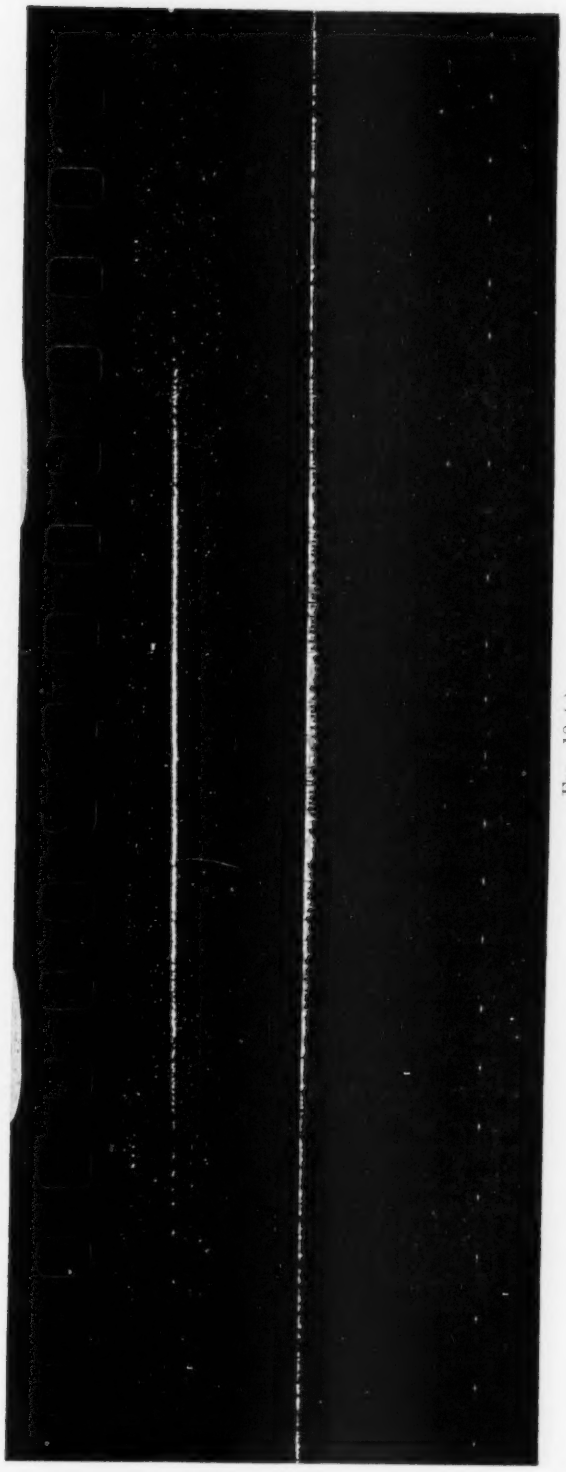


FIG. 18 (*c*)
Photograph of radar reflection signals from Sputnik I, taken at Jodrell Bank, Cheshire.

orbit data have been forthcoming. Figure 18b illustrates a photograph of the track of Sputnik II passing over Edinburgh.

Radio observation of the track of the Sputnik III satellite has been possible throughout its career as it was powered by solar batteries. The rocket case was a bright object, reaching a magnitude -2 under favourable conditions. Great fluctuations of brightness occurred, by as much as a factor of 100 on occasion, presumably due to rotation. The period of the fluctuations changed gradually from 8·5 sec in mid-July to 9·5 sec at the end of October. Apart from being less bright than the rocket case, the instrumented satellite behaved optically in a very irregular fashion, sometimes being invisible and at others flashing out at magnitude $+1$.

The Explorer satellites contain two radio transmitters both of which transmit telemetry as well as tracking information; the higher power one (108·03 Mc/sec, 60 mW) is designed for a lifetime of approximately 1 month and the lower power one (108 Mc/sec, 10 mW) for about twice as long. Accurate radiotracking of these satellites is carried out either by the Minitrack or Microlock systems. The low-power transmitter in Explorer I ceased operating on 1958 May 23 and the high-power one about the end of February 1958. (See: Jet Propulsion Lab. Ext. Publication No. 461; Pasadena, 1958 February 28.) In Explorer III the corresponding dates were 1958 June 14 and May 11, respectively.

The Vanguard Test Sphere

Although the launching of the small test sphere in the Vanguard project was essentially part of the pre-launching tests for the main programme, in many ways it represents a remarkable achievement. Thus it must be remembered that not only the 3 lb sphere but also the 50 lb rocket case were placed in an orbit with perigee height as great as 640 km. At such great altitudes the effect of air resistance is so small that the sphere is likely to remain in orbit for 100 years or more before plunging into the atmosphere. In addition to this the sphere carries two radio transmitters, one at 108 Mc/sec powered by mercury alkaline batteries at a level of 10mW, and the other at 108·03 Mc/sec which, for the first time, uses solar batteries as a power source. Four solar cells are fitted so that when in sunlight at least one is in operation. The mean power level is about 5mW. Up to the time of writing these solar batteries have been operating satisfactorily and it is a matter of great interest and importance to determine their lifetime, which will certainly depend strongly on the rate at which the sensitive surfaces are modified by micrometeoritic bombardment.

It is possible to obtain the temperature of the transmitters from accurate measurements of their frequencies, so this is a further source of useful data.

Effect of Air Drag on the Orbits

The most obvious way in which the effect of air drag is manifest is by the change in period of revolution of the satellite with time. Figure 19 illustrates the variation of period with number of revolutions for the rocket case of Sputnik I. This includes data obtained by radar as well as by optical means.

The last observation of the rocket case was made from California at 00.12 hr U.T. on 1957 December 1 and it seems that it could not have survived much longer. Its probable point of descent into the atmosphere lies between California and Cape Horn.

The behaviour of Sputnik II is rather less regular than that of Sputnik I. This is made particularly clear if the observed rate of decrease of period is plotted against

time as illustrated in Fig. 20. Very marked irregularities are noted which exceed the quite considerable observational errors.

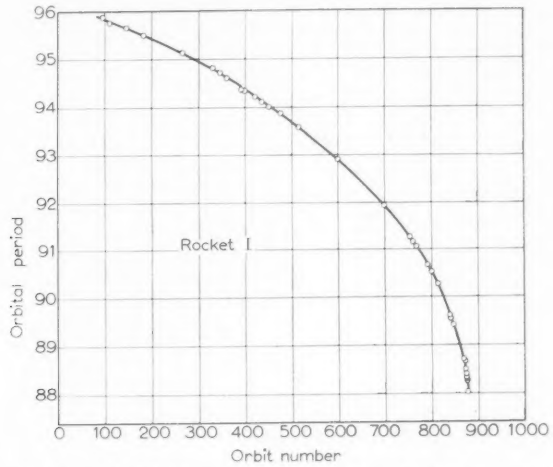


FIG. 19. Observed variation of orbital period with number of orbital revolutions for the rocket case of Sputnik I. The observed points include optical, radio and radar observations. This curve was kindly supplied by Prof. A. C. B. LOVELL, F.R.S.

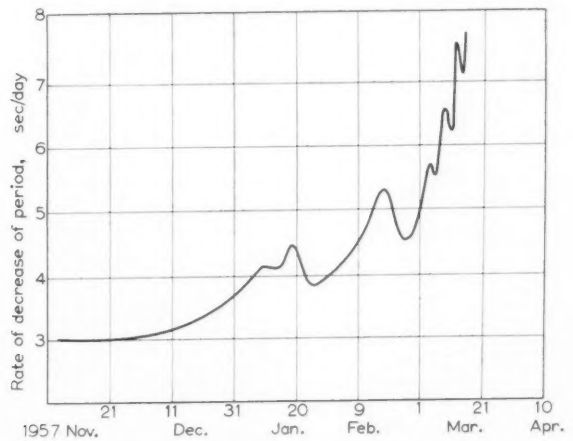


FIG. 20. Observed variation of rate of change of orbital period with time for Sputnik II.

Because of the irregularities it was difficult to predict when Sputnik II would descend into the atmosphere but it seems clear that this occurred between 1.50 and 2.00 hr U.T. somewhere over the West Indies or the North of South America. A brightly glowing object was observed in that region at the expected time.

The irregularities observed in the day-to-day variation of period of revolution with time for Sputnik II were also very marked both for the Sputnik III instrument carrier and rocket case (see Fig. 21). The behaviour of the rocket case became more irregular towards the end of its life (1958 December 3). The instrument carrier is likely to circulate until some time in 1960.

For the Explorer and Vanguard satellites the rate of change of period is less irregular but nevertheless exhibits some surprising features. This may be seen in

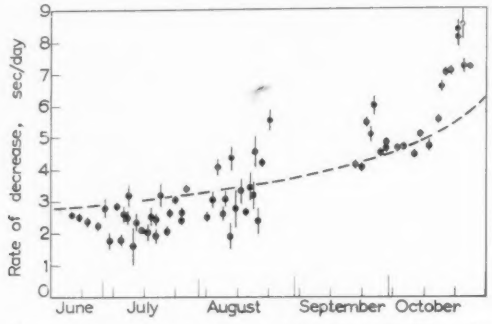


FIG. 21. Rate of change of period for Sputnik III rocket case (1958, June to October).

Fig. 22 for Explorer I and Vanguard (test sphere). From the time of launching up till the end of August the period of revolution for Explorer I decreased linearly with

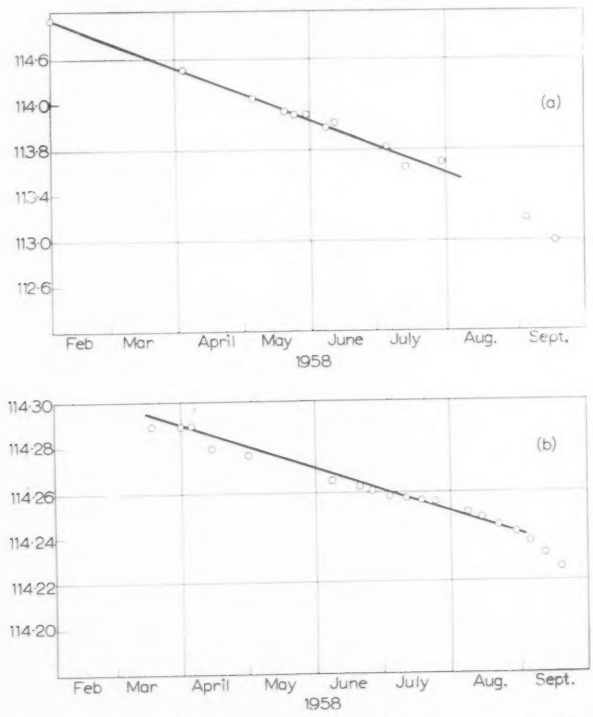


FIG. 22. Change of period (given in minutes) with time
(*a*) Explorer I
(*b*) Vanguard test sphere.

time (apart from quite small fluctuations). A sudden increase then occurred in the rate of change of period and this has been maintained. A similar result is noted at about the same time for the Vanguard sphere.

Abrupt changes in the slope of the period–time curve are also revealed in the behaviour of Sputnik III instrument carrier and rocket case, if the irregular fluctuations are averaged out. Thus for the rocket case the rate of change of period was about 2·4 sec/day from 20 June to 20 July, 3·2 sec/day from 31 July to 19 August, 4·0 sec/day from 23 August to 22 September, and near 4·8 sec/day from 24 September to 10 October. It may be significant that the average time between these abrupt changes is nearly 27 days.

Determination of Air Density

The problem of determining the air density near the perigee altitude from the observed variation of time rate of change of period is complicated by two factors. One is the difficulty of deciding what is the correct value to take for the drag coefficient when the object is not spherical in shape and will be rotating in a complicated manner. The second concerns the interpretation of the irregularities in the period time variation which is a marked feature of the observations for several objects.

Considering the latter problem first, the most important question to answer is whether the irregularities really arise from changes in air density at perigee or from sudden changes in drag coefficient due to changes in the mode of rotation of the object. Two pieces of evidence indicate strongly that the effect cannot be ascribed to sudden changes in rotational mode. The Vanguard satellite is spherical (apart from the radio antennae) and yet shows an abrupt change. The strong fluctuations of brightness for the Sputnik III objects are very likely to be due to rotation. Certainly for the rocket case there were no sudden changes of period of the fluctuations at least up to mid-October, suggesting that no abrupt change of rotational mode occurred. In view of these considerations we must assume that the irregularities do actually arise from fluctuations in air density at the perigee.

This is not so surprising when account is taken of the air density measurements carried out at Fort Churchill using vertical sounding rockets (LAGOW, HOROWITZ and AINSWORTH, 1958). These observations showed that at altitudes above 100 km the density varies markedly with time of day, season and latitude.

It has been suggested that, as the intervals between the abrupt changes of slope of the period–time curve for the Sputnik III objects are nearly 27 days, there might be some correlation with solar disturbances. These show a 27 day recurrence tendency because of the rotation of the Sun. NONWEILER (1958) has attempted to trace a correlation in some detail between the solar flare intensity (measured by the product of area and duration) and the rate of decrease of period for Sputnik II. Some evidence of a correlation exists but if so it requires that a solar flare is associated with a decrease in air density on a world-wide scale. Further indication of solar control has been provided by the correlation found between the fluctuations of period and the intensity variation of the 20 cm radiation from the Sun (PRIESTER 1959) for Sputnik III, and with that of the 10·7 cm radiation (JACCHIA 1959) for Vanguard.

Many investigations have derived data about the air density by applying formulae such as (3.51). GROVES (1959) has carried out an analysis using a generalization of this formula which allows for a variation of scale height with altitude. He takes the variation of air density with height above a standard altitude h_0 to be of the form

$$\log \rho/\rho_0 = -\{(h-h_0)/\mathrm{H}_1\} + (h_0-h_1)^2/H_2^2. \qquad \ldots . (13.1)$$

This has the effect of modifying the expression for α_2 and ϵ in (3.31) but preserves

the form of the relation unaltered. For the Explorer satellites the drag coefficient was taken as given by SCHILLING (1958) who derived it on the assumption that the cross section to be taken is the mean over all orientations, supposed equally likely. Allowance is also made for the contribution to the drag from the radio antennae. For the Vanguard test sphere this allowance is the only source of uncertainty as it is also for the Sputnik I satellite. Owing to the relatively inaccurate observations of perigee height for both the Sputnik I objects, the data from them were mainly used to estimate the drag coefficient for Sputnik II for which much more reliable orbit data are available. The procedure was to obtain the drag coefficient for the Sputnik I rocket case from the observed result that its initial rate of change of period was 1·84 times faster than for the satellite sphere. The same value was then assumed for Sputnik II.

For each set of satellite data, perigee air density was calculated using assumed values for H_1 and H_2. The relation (13.1) was then fitted by the method of least squares to nine such values, after suitable weighting, so giving H_1, H_2 and log ρ_0 corresponding to an arbitrary h_0 of 200 km. The calculation was repeated until a sufficiently convergent solution had been reached, H_1 being obtained as 46($\pm$ 5) km and $(H_1/H_2)^2$ as 0·028 ($\pm$ 0·013). Table 4 gives the values obtained for the air density between 150 and 700 km.

*Table 4**

Air density derived from analysis of satellite observations

Altitude (km)	*Density* (g/cm^3)	*Altitude* (km)	*Density* (g/cm^3)
150	$1{\cdot}2 \times 10^{-12}$	450	$3{\cdot}8 \times 10^{-15}$
200	$3{\cdot}8 \times 10^{-13}$	500	$1{\cdot}9 \times 10^{-15}$
250	$1{\cdot}4 \times 10^{-13}$	550	$9{\cdot}6 \times 10^{-16}$
300	$5{\cdot}0 \times 10^{-14}$	600	$5{\cdot}3 \times 10^{-16}$
350	$2{\cdot}0 \times 10^{-14}$	650	$3{\cdot}2 \times 10^{-16}$
400	$8{\cdot}4 \times 10^{-15}$	700	$2{\cdot}0 \times 10^{-16}$

It must be remembered that these values represent averages in which the weighting depends on the motion of the perigees of the different satellites. Other investigators (SCHILLING and STERNE 1958, KING-HELE 1959a) have derived air densities which are generally similar to those obtained by GROVES but differ in detail. At 200 km altitude rocket-observations have given the following densities in g/cm^3 (LAGOW, HOROWITZ and AINSWORTH 1958). $(1{\cdot}4 \pm 0{\cdot}5) \times 10^{-13}$ (33° N, summer day), $(6{\cdot}7 \pm 0{\cdot}2) \times 10^{-13}$ (59° N, summer day), $(3{\cdot}6 \pm 1{\cdot}5) \times 10^{-13}$ (59° N, winter day), $(1{\cdot}3 \pm 0{\cdot}6) \times 10^{-13}$ (59° N, winter night). In view of the variability of the density, there does not seem to be any serious inconsistency with the value given in Table 4. However, it is worth noting that there does not seem to be any evidence of such marked seasonal and latitude effects from the satellite data (KING-HELE, 1959b) which suggest that these effects, in the altitude range 200–250 km, do not produce departures from the average value by more than a factor of 1·5.

A direct measurement at an altitude of 368 km was made using an ionization gauge of the type described on p. 67 carried in Sputnik III. The value found,

* A recent analysis by KING-HELE (1959b), including data from fifteen satellite objects, suggests that the values given in Table 4 should be increased at the higher altitudes to 2·6, 0·87 and $0{\cdot}31 \times 10^{-15}$ g/cm^3 at altitudes of 500, 600 and 700 km, respectively, but considerable uncertainty still remains in these values.

$9{\cdot}2 \times 10^{-15}$ g/cm^3, is not inconsistent with those of Table 4 when allowance is made for the difficulty of the measurement and the variability of the density.

The scale height at 200 km of 46($\pm$ 5) km obtained by GROVES is to be compared with the summer values obtained by rocket observations (LAGOW, HOROWITZ and AINSWORTH, 1958) which are 93 km at 59° N and 38 km at 33° N. GROVES has also examined the possibility of determining the scale height by using the formula (3·47) generalized to include the quadratic term in the assumed density formula (13·1) and some sixty observations on the rate of change of period of Sputnik II. The scale height found was 36($\pm$ 15) km at 200 km altitude, the low accuracy arising from the rather large fluctuations in the rate of change of period.

The main feature of the air density structure of Table 4 is the relatively great extension of the atmosphere which it indicates, requiring a considerably higher temperature than expected for the outer atmosphere. Some of the problems which this raises have been discussed by BATES (1959).

Effect of the Figure of the Earth on the Orbits

These effects have been discussed on p. 55 the main ones being the rotation $\mathrm{d}\Omega/\mathrm{d}t$ of the plane of the orbit about the Earth's axis (the regression of the nodes) and the rotation $\mathrm{d}\omega/\mathrm{d}t$ of the major axis in the orbital plane (rotation of the perigee). Theoretical formulae for $\mathrm{d}\Omega/\mathrm{d}t$ and $\mathrm{d}\omega/\mathrm{d}t$ are given in (3.62 and 3.65) and (3.64) respectively. Table 5 gives a comparison between the observed and calculated values of $\mathrm{d}\Omega/\mathrm{d}t$ and $\mathrm{d}\omega/\mathrm{d}t$ (KING-HELE and MERSON, 1958). For Sputnik I the observations were made at the radio interferometer station of the Royal Aircraft Establishment (RAE) at Lasham while for Sputnik II they were made optically by kinetheodolites at the RAE stations at Orfordness, Aberporth and Bramshott. The numerical values assumed for J and D in (3.65) and (3.64) were as given on p. 53.

Table 5

Comparison of observed and calculated values of $\mathrm{d}\Omega/\mathrm{d}t$ *and* $\mathrm{d}\omega/\mathrm{d}t$ *for Russian satellites*

Object			Sputnik I	Sputnik II		
Date			18·5 Oct. 1957	25·0 Nov. 1957	4·0 Jan. 1958	13·0 Feb. 1958
$\frac{\mathrm{d}\Omega}{\mathrm{d}t}$	obs.	°/day	3·17 $\pm$0·03	2·685$\pm$0·003	2·814$\pm$0·003	2·992$\pm$0·003
	calc.		3·176$\pm$0·024	2·705$\pm$0·003	2·832$\pm$0·003	3·009$\pm$0·003
$\frac{\mathrm{d}\omega}{\mathrm{d}t}$	obs.		—	−0·403$\pm$0·014	−0·422$\pm$0·003	−0·447$\pm$0·004
	calc.		−0·40 $\pm$0·05	−0·409$\pm$0·017	−0·428$\pm$0·017	−0·455$\pm$0·017

It will be seen that small but significant differences exist between observed and calculated values of $\mathrm{d}\Omega/\mathrm{d}t$ for Sputnik II.

In fact the ratio of the theoretical to the observed value of $\mathrm{d}\Omega/\mathrm{d}t$ remained effectively constant at the value 1·0068 $\pm$ 0·0003 throughout the life of Sputnik II. This excludes any explanation based on atmospheric effects, which should increase steadily with time. It also makes it unlikely that the discrepancy is due to neglect of a third harmonic in the expansion (3.53) of the Earth's gravitational potential. According to (3.70) this would give an effect proportional to $e\sin\omega$ which decreases markedly during the lifetime of the satellite. We are therefore forced to seek

an explanation in a modification of the numerical values of the quantities J and D used in obtaining the theoretical results. To decide which needs modification it is necessary to turn to data obtained by observation of satellites circulating in orbits of markedly different inclination. Such data are available from observations of the Explorer and Vanguard satellites which show that for these satellites also the theoretical values of $\mathrm{d}\Omega/\mathrm{d}t$ exceed the observed. As the effect of D changes sign in going from an inclination of 65° to one of 35°, whereas that of J remains of the same sign, it is clear that the major correction must concern the latter quantity.

King-Hele (1959a) thus derived the values

$$J = 1{\cdot}6245 \pm 0{\cdot}0003 \times 10^{-3}$$
$$D = 5{\cdot}7 \pm 0{\cdot}9 \times 10^{-6}$$

giving for the ellipticity α the value $(298{\cdot}21 \pm 0{\cdot}03)^{-1}$ in place of the original value $(297{\cdot}1)^{-1}$. In obtaining these results King-Hele allowed for the small effects due to a third harmonic (see below), to the sideways force caused by the rotation of the atmosphere, and to solar and lunar perturbations. Applying formulae derived by Kozai (1959), it is found that the Moon increases $\mathrm{d}\Omega/\mathrm{d}t$ by about 1 part in 10,000 for Vanguard I and 1 part in 30,000 for Sputnik II. The Sun is about one half as effective.

Jacchia (1958), using the same data for Sputnik II as does King-Hele, combined with precise data for the Vanguard satellite, gives results which agree with those of King-Hele. Results obtained from observations of Explorer I and Vanguard satellites by O'Keefe, Hertz and Marchant (1958) are also in agreement.

The contribution ϵ_3 from the third harmonic (see (3.70)) has been derived by O'Keefe, Eckels and Squires (1959) and Kozai (1959) from the perturbation of the eccentricity of the orbit of the Vanguard satellites with a period equal to that of the revolution of the perigee. They find

$$\epsilon_3/GMR^3 = -2{\cdot}2 \pm 0{\cdot}1 \times 10^{-6}$$

thus providing the first definite evidence of a difference between northern and southern hemispheres.

Secular Change of Orbital Inclination

King-Hele and Merson (1958) found from analysis of the kinetheodolite observations of Sputnik II that the inclination of the orbit changed from its initial value of 65·32° to about 65·19° at the end of its lifetime. A similar effect was found for the Sputnik III rocket case, the change being from 65·18° initially to about 65·08° at the end of October 1958; (King-Hele 1959a).

This effect must be due to a perturbing force with a component in the plane of the orbit along the normal. The fact that the decrease of inclination is much more rapid near the end of a satellite's lifetime points to an atmospheric source of the disturbance. One possibility is the effect of the rotation of the atmosphere which will apply a perturbing force of the required type, tending to decrease the inclination for a satellite travelling from West to East. Plimmer (1958) has calculated the change of inclination per revolution due to this effect. He finds that

$$\Delta i = \frac{\sin i}{3} \frac{[1-(15/8)k-2e]\cos^2\omega + 2k}{(1-\tau\cos i)[1+(k/8)]+2e}\{1+O(k^2, e^2)\}\Delta\tau \qquad \ldots . (13.2)$$

where $k = H/ae$, H being the scale height, and $\Delta\tau$ is the change of period τ per revolution.

Comparison of (13·2) with observation, assuming a scale height of 54 km, reveals quite good general agreement but this would be improved if the theoretical variation were more rapid. This could be explained if, in addition to the rotation, there was a strong prevailing West-to-East wind. More data are required before other explanations can be excluded. It may be that the rate of change of inclination of satellite orbits will give valuable information about wind speeds at high altitudes.

Radio Observations of the Russian Satellites and the Ionosphere

The choice of frequencies for the radio transmitters in the Russian satellites opened the prospect of obtaining new information about the ionosphere by use of the methods outlined on p. 83. Clear indication of the marked influence of the ionosphere on the propagation, particularly of the 20 Mc/sec signals, is provided by the fact that, on many occasions, these signals were received when the satellite was well below the horizon and many other examples of complicated propagation paths were noted.

Furthermore the Doppler records exhibit discontinuities of several tens of cycles in the Doppler frequency shift which are of irregular duration and occur at random. In many cases also a time difference of several seconds was observed between the instants of closest approach of the satellite indicated by the signal tracing and Doppler shift. KITCHEN and JOY (1958) have shown how the latter effect may arise if the lower edge of the F-layer is tilted. Details of these studies are reported in papers by KITCHEN (1958), BAIN (1958), BOOTH (1958), KITCHEN, JOY and BELLAM (1958), BAIN and SHEARMAN (1958) and WARWICK (1958).

Analysis of the radio data to derive ionospheric information has been hampered by the lack of the systematic accurate optical observations necessary to determine true positions at different times. Enough has been achieved already, however, to show that much can be expected from future satellites for which precise tracking has been well organized.

One of the most promising methods of analysis is that which interprets fading of the signals in terms of the Faraday effect. The fact that signals were transmitted at two frequencies greatly assisted this analysis. Apart from the Faraday effect as discussed on p. 84, fading could arise in two ways from rotation of the satellite. On the one hand, this could cause intensity variation due to rotation of the polar diagram of the transmitting aerial. Further, any plane-polarized component emitted, would produce a variable signal in a linearly-polarized receiver aerial. Both of these effects could produce variations with a frequency either equal to or twice that of the satellite's rotation, but they may be distinguished by comparing the signals on two aerials which have perpendicular planes of polarization. It was found that these signals differed in phase by $\pi/2$ showing that rotation of the plane of polarization either by satellite spin or the Faraday effect was responsible. These two may then be distinguished by the fact that the Faraday effect is proportional to the square of the wavelength whereas the effect of satellite spin is independent of wavelength. In this way, by observing at 20 and 40 Mc/sec and also on the second harmonic of the latter, it was found that the satellite sphere was spinning at 7 r.p.m. and data on the Faraday effect were obtained.

The analysis of these data, and also that derived from the Doppler frequency records for the two transmission frequencies on the lines indicated on p. 83, is still incomplete but indicates (WEEKES, 1958) that the rate of decrease of electron con-

centration in the first 50 km above the F_2-layer maximum is very slow indeed, a result which is in accordance with current ideas about the origin of the layer.

A similar conclusion is derived by AL'PERT, CHUDESENKO and SHAPIRO (1958) from an analysis of the times of "radio rise" and "radio set" of Sputnik I, the times of rising and setting as judged from the onset or disappearance of radio signals from the satellite. Attention was limited to those cases in which these times were sharp and complications due to "multiple-hop" propagation were absent. The work was directed towards obtaining information about the rate of fall of electron concentrations above the F_2 maximum. It was assumed that, at a height h above that h_m of the maximum, the electron concentration was $\exp[-k(h-h_m)]$ times the maximum value, k being a constant to be determined. The layer below the maximum was assumed, as usual, to be parabolic in form. The critical frequency, the height h_m and that h_0 of the lower edge of the layer were all obtained from ionospheric sounding records and the orbit of the satellite from optical observations. It was found that the range of values of k required to fit the observations on different occasions was from 2 to 7×10^{-3} km^{-1}, the most probable value being $3{\cdot}5 \times 10^{-3}$ km^{-1}. This corresponds to a scale height of 300 km, a very gradual fall off in electron concentration. The assumption of a simple exponential form for the variation of electron concentrations above the maximum is rather crude but it is, nevertheless, clear that the layer is very thick.

Experiments in situ *carried out from Russian Satellites*

As far as is known at the time of writing the satellite sphere of Sputnik I contained no instruments apart from the radio transmitters. Sputnik II on the other hand contained Geiger counters for measuring total cosmic ray intensity, as well as instruments for measuring temperature and intensity of radiation. In addition there was also the elaborate chamber which housed the dog and which necessarily included apparatus for ensuring supply of oxygen, food and water as well as instruments for observing the heart beats, blood pressure, etc. of the animal. No reports are yet available as to the data obtained although it is understood that, during the week in which the battery power for data transmission was available, many useful results were forthcoming. Thus the cosmic ray intensity transmitted showed the expected geomagnetic control, indicating that the apparatus was functioning correctly.

Sputnik III carried a great variety of instruments. These included magnetometers, counter equipment for the measurement of cosmic rays and other radiation, ionization gauges, a screened probe for measurement of positive ion concentration and satellite potential (as described on p. 84), a radio frequency mass-spectrometer for determination of the nature of the positive ions in the ionosphere, a type of electrostatic voltmeter for the measurement of electric field strength in the ionosphere and impact microphones for the detection of micrometeorites.

Many of these instruments seem to have been working quite satisfactorily and preliminary reports have been given of some of the results obtained. The atmospheric density measurement reported has been referred to on p. 98.

The technique described on p. 84 showed that, on the first day of flight at an altitude of 795 km, the positive ion concentration was $1{\cdot}8 \times 10^5$/cm^3 and the satellite was charged to a potential 6·4 V below its surroundings, while at an altitude of 200 km the corresponding values were $4{\cdot}2 \times 10^5$/cm^3 and 2 V respectively. Observations at night gave satellite potentials of 1 V or less below the surroundings.

The mass spectra obtained in the first few circuits (ISTOMIN 1958) showed that O^+ ions predominate at altitudes of 250–950 km. N^+ ions were also found in proportion to the O^+ ions increasing from 0·037 at 250 km altitude to 0·06 at 820 km.

Evidence was obtained from the magnetometer measurements made in Sputnik III of the existence of local ionospheric current systems. This was shown from the occurrence of negative and positive peaks in the variation of the magnetic field. A very interesting result was also found from the magnetic measurements made in the first Russian deep space probe, launched on 1959 January 22, which became an artificial planet. The gradual decrease of field with distance from the Earth persisted out to about 20,000 km but over the next 4000 km it showed a marked anomaly. Thus the field actually increased to a maximum at 22,000 km before decreasing again. This is specially significant in view of the suggested existence of a ring current around the Earth set up from solar corpuscular streams (see p. 75).

Very interesting results have been obtained from the measurement of radiation intensities with counters in Sputnik III and in the Russian deep space probes. We shall defer a description of these results to p. 105 when they will be discussed in relation to the whole investigation of the great radiation belts.

A Čerenkov-type of discriminating counter similar to that described on p. 73 for observation of the intensity of the heavy component of cosmic rays operated successfully. Preliminary results obtained over a period of ten days gave $1{\cdot}2 \pm 0{\cdot}1$ nuclei in the primary radiation encountered per minute with atomic number $Z > 16$ but in the whole period of 9 days there was only one record which might have been due to a nucleus with $Z > 30$.

The apparatus for micrometeorite detection consisted of a piezo-electric pickup, of ammonium phosphate, which responded to the impact of the dust particles on a plate. This device recorded, at least according to the preliminary records, 90 impacts per m^2/sec (NAZAROVA 1958). It is still not certain whether the pickup response is determined by the energy or the momentum of the incident particle, so that until this point is settled it is not yet possible to derive reliable information about the momentum or energy distribution of the micrometeorites encountered.

Experiments in situ *carried out from American Satellites*

The first two Explorer satellites contained a number of instruments. In each there was a Geiger counter to measure the total cosmic ray intensity as well as thermistors to measure skin and interior temperature, impact microphones and erosion gauges for study of micrometeorites.

In Explorer I, the first American satellite to be launched, there was continuous emission of telemetry information requiring the maximum number of receiving stations distributed over a wide longitude range within the latitude range covered by the orbit. In addition to continuous transmission, data storage was provided in Explorer III from which information was only transmitted on command. Useful opportunities of checking the operation of this system were provided by comparison with results obtained from the continuous systems in both satellites.

Accurate radiotracking of these satellites was carried out either by the Minitrack or Microlock systems.

Very interesting indications about radiation intensity were obtained from the Geiger counters in Explorer I and III. In each case the single counters had stainless steel walls 0·05 in. thick and were surrounded by the stainless steel cases of the

payload of 0·023 in. thickness so that there was about 1·5 g/cm^2 of stainless steel absorber. It was found that for altitudes below about 1200 km the counting rates for both satellites were normal. The variation with height over California was observed from 400 to 1400 km and showed an approximately linear increase with height which, if extrapolated to 100 km, agrees quite well with the counting rate observed for cosmic rays in rocket flights. At much higher altitudes, however, a remarkable effect was observed. Above 2000 km no counts were recorded over periods of 2 min or so whereas at the altitudes below 1000 km there were about 30 counts/sec. It was inconceivable that the cosmic ray intensity could have fallen to such very low values. Instead, the effect seemed to be due to blocking of the counters by a very high intensity of incoming radiation. This will occur if the interval between arrival of particles at the counter is short compared with the dead time of the counter. Laboratory experiments using X-rays showed that this blocking would indeed occur when the true counting rate reaches about 35,000 counts/sec. It seemed probable that true counting rates at altitudes of 2000 km or more did reach these values. Partial confirmation of this explanation was afforded by the data derived on command from Explorer III. This provided information over continuously varying ranges of height and the transition from the normal to the blocking condition was found to involve passage through a region in which the counting rate rose rapidly.

In order to verify this interpretation and to explore the region of intense radiation more thoroughly VAN ALLEN took advantage of the launching of Explorer IV to incorporate more elaborate counters in that vehicle. Counters *A* and *B* were circular disc scintillators, the former being of plastic and the latter of sodium iodide. Both were mounted with the disc near an open hole in the payload shell so as to face in opposite directions. *C* and *D* were Geiger counters located side by side on the axis of the satellite. *C* was not deliberately shielded but *D* was surrounded by a lead cylinder of 1·6 g/cm^2 thickness.

The results obtained (VAN ALLEN *et al.*, 1958) confirmed the conclusions and extended the results obtained with the earlier satellites. At a geographical longitude of 80 ± 20° W and low latitudes the intensity recorded by all counters increased slowly to altitudes of 1000 km and then very rapidly up to the highest altitude reached, 2200 km.

Figure 23 illustrates a plot of the contours of constant counting rate in the meridian section at this longitude. It will be seen that the contours are approximately symmetrical about a latitude of 14° S which is close to the magnetic dip equator. This

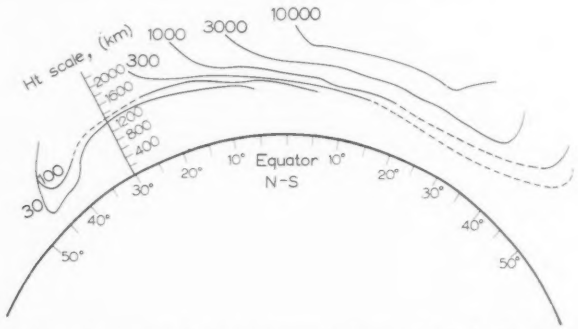


FIG. 23. Contours of constant counting rate. Data for Explorer IV. Longitude 80° ± 20° W.

association with geomagnetic co-ordinates is also exhibited by the contours for the longitude 140° ± 20° E and strongly suggests that the observed radiation has built up in some way through trapping of solar corpuscular streams by the Earth's magnetic field. Some supporting evidence is provided from data obtained with Explorer I. These showed a considerable decrease in the heights at which the intense radiation was encountered after the brilliant auroral display of 1958 February 10.

Further Study of the Radiation Belt

A remarkable amount of further information about the field of radiation surrounding the Earth has been obtained in the year or so since the first observations made by VAN ALLEN which are described above. (For a general reference see SINGER, 1959.)

The cosmic ray detecting equipment in Sputnik II already showed some indication of unexpectedly large intensities. Thus on 1957 November 7, when passing over a region at 55° geomagnetic latitude, the instruments recorded a 50 per cent increase in radiation intensity (VERNOV, 1958) which could not be attributed to an increase in intensity of the cosmic radiation. Much more detailed information has been obtained from equipment in Sputnik III (KRASSOVSKII, 1959). This consisted of two fluorescent screens covered with aluminium foils of different thickness, the surface densities being 4×10^{-4} and 8×10^{-4} g/cm^{-2}. Electrons, positive ions and X-rays, with energy greater than a few MeV, impinging on the screen produce radiation the intensity of which can be measured by an arrangement of photomultipliers. Strong signals were recorded which were probably due to incidence of electrons, with about 10 MeV energy, on the foils. The intensity of the signals increased with distance from the Earth and was greater in polar than in equatorial regions.

In addition to the fluorescent screen detectors a scintillation counter was also included and shielded so that it responded only to particles of much higher energy. Signals due to protons of mean energy about 100 MeV were obtained with considerable frequency at high altitudes and low latitudes.

The rather obscure situation about the nature and extent of the radiation zones was greatly clarified by the subsequent launching of the American and Russian deep space probes, all of which contained instruments for observing particle radiation. It was found first by VAN ALLEN (1959) that there are two radiation zones, an inner and an outer, a result confirmed and extended by the observations of VERNOV and CHUDAKOV (1959). VAN ALLEN's counters were carried in Pioneer III which reached a distance of 107,400 km from the Earth. Measurements of radiation intensity were made, both on the outgoing and return path of the vehicle, out to nearly 8 Earth radii. They showed that, in the geomagnetic equatorial plane, there is an inner zone in which the intensity is a maximum at about 1 Earth radius distance. There follows a region of relatively low intensity until a second zone begins to develop in which the minimum occurs at a distance of about $3\frac{1}{2}$ Earth radii.

VERNOV and CHUDAKOV were able to show, from a pulse height analysis of the response of the scintillation counter included in the first Russian artificial planet, not only that two zones exist but that the particles in each have different energy distributions. In the inner zone, as indicated from observations made with Sputnik III there are protons with mean energy of 100 MeV as well as electrons with much

smaller mean energy. By contrast there are very few energetic protons in the outer zone, most of the observed particles being electrons of mean energy 25 keV.

Combination of all the observations leads to a picture of the radiation belts which is very closely related to the Earth's magnetic field, as shown in Fig. 24. There are insufficient data to establish that this picture is fully correct but all the results up to the present are consistent with it.

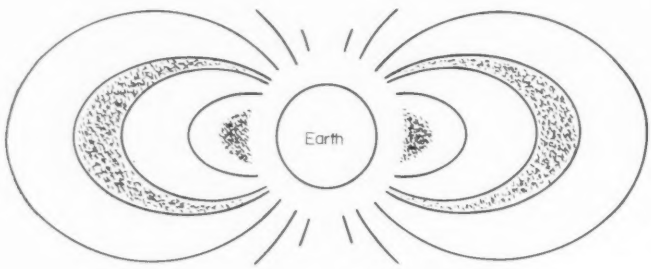


FIG. 24. Illustrating the position of the inner and outer radiation belts (shown shaded) in relation to the Earth and the Earth's magnetic field, lines of force of which are shown.

The close association with the Earth's magnetic field naturally directs attention to STÖRMER's calculations (STÖRMER 1913) on the orbits of charged particles in the field of a magnetic dipole. He found that, under certain circumstances, particles could be trapped in orbits between two conjugate points on a line of force, disposed at equal distances from the magnetic equator. If charged particles are trapped in this way for long periods then a considerable concentration will build up, provided there is some source which makes up for the slow but finite loss due to collisions, magnetic field fluctuations, etc. There has been extensive discussion about the sources which maintain the two zones, but although the situation is not completely certain, it seems likely that the inner zone is maintained from the Earth's cosmic ray albedo and the outer from solar corpuscular streams. The supporting evidence is as follows.

Primary cosmic rays reacting with atmospheric nuclei will produce secondary neutrons some of which will move outwards, unrestricted by the Earth's magnetic field. Eventually, however, they will decay into protons and electrons which can be trapped by the Earth's field. As the neutrons will possess energies of a few hundred MeV on the average, we can expect the protons to have energies of the same order, the electrons about 2000 times less. This is consistent with the data on the energy distribution of the particles in the inner zone obtained by VERNOV and CHUDAKOV. A further experiment was carried out by FREDEN and WHITE (1959) in which nuclear sensitive emulsions were exposed at an altitude of 1200 km from a vertical sounding rocket, recovered and developed. From the tracks obtained they were able to confirm that the energetic particles in the inner zone are protons with energies up to 900 MeV. Quantitative calculations of the expected intensities are very difficult and it will be some time before this may be done with sufficient confidence to provide a valuable check on the hypothesis.

The solar origin of the outer belt is consistent with the fact that hardly any energetic protons are found there, the main particles being electrons of 25 keV energy. Further support for the suggested origin is provided from observations

made by van Allen (1959b) from instruments carried aboard Pioneer IV, the first artificial planet launched by the U.S.A. It so happened that, during the five days prior to launching, there was marked solar activity. The intensity in the outer belt was found to be several times greater than during the flight of Pioneer III. This is what would be expected if there had been intense injection of charged particles from the Sun.

Confirmation of the occurrence of trapping was provided by the artificial production of a radiation belt in the "Argus" experiment (Christofilos 1959). Kiloton atomic bombs, carried aloft by high-flying rockets, were exploded at altitudes of about 480 km above the South Atlantic. Electrons resulting from the fission products were trapped in orbits which gradually precessed westwards round the Earth, as predicted, to form a radiation belt. This was confirmed (van Allen, McIlwain and Ludwig 1959c) by observations made with the counter equipment carried in Explorer IV which circulated through the region in which the belt was produced.

Three experiments on these lines were carried out, the explosions taking place at locations 38° S, 12° W; 50° S, 8° W; and 50° S, 10° W respectively. In all cases a radiation belt over 90 km thick was detected for a period of five days or so after the explosion. During these periods there was no appreciable disturbance of the natural inner zone—the artificial belts fell in the region between the two natural zones.

Although there has been remarkably rapid progress in observation and interpretation of the nature and region of the radiation belts, a great deal remains to be done. This applies particularly to such questions as the detailed morphology of the belts, the variation of the belts with time and the mechanism of trapping and of loss of particles from trapped orbits. Correlated systematic observations of radiation intensity and magnetic field are likely to be very helpful, using instruments carried in satellite vehicles circulating in highly eccentric orbits.

The relation of the outer belt to geomagnetic disturbances and auroral phenomena is obviously a matter of much interest. It is surely significant that the outer zone approaches the Earth most closely at the auroral zones, and it is not difficult to imagine that, when the Sun injects many more particles into the zone, some particles are spilled out into the atmosphere. If sufficiently energetic they will produce auroral displays. However, the mechanism of acceleration, which seems to be necessary, is as unclear as in presatellite days. The relation to the different phases of a magnetic storm is also obscure. Further exploration of the space around the Earth out to a few Earth diameters will surely help to build up a more detailed picture of solar-terrestrial relations than ever before.

Acknowledgements.—I am much indebted to Mr. D. G. King-Hele, Dr. G. V. Groves and Dr. K. Weekes for providing me with the results of their investigations prior to publication, and to Sir Bernard Lovell, F.R.S. and Dr. H. E. Butler for assistance in obtaining illustrations. Dr. G. V. Groves has also afforded me invaluable assistance by reading and checking the manuscript and proofs. My niece, Mrs. M. Farley, helped in the preparation of the manuscript at a stage when I was in Australia.

REFERENCES

Al'pert, Y. A., Chudesenko, E. F. and Shapiro, B. S.	1958	Preliminary results of Scientific Research on the First Soviet Artificial Earth Satellites and Rockets. Acad. Sci. U.S.S.R. IGY Programme, 11, 40.
Bain, W. C.	1958	*Proc. Roy. Soc. A.* **248,** 68.
Bain, W. C. and Shearman, E. D. R.	1958	Report to IGY Symposium on Rockets and Satellites, Moscow.
Bates, D. R.	1959	*Proc. Roy. Soc. A.* **253,** 450.
Blitzer, L., Weisfeld, M. and Wheelon, A. D.	1956	*J. appl. Phys.* **27,** 1141.
Blitzer, L.	1957	*Jet Propulsion* **27,** 405.
Blitzer, L. and Wheelon, A. D.	1957	*J. appl. Phys.* **28,** 279.
Booth, C. F.	1958	*Proc. Roy. Soc. A.* **248,** 72.
Chapman, S. and Bartels, J.	1940	*Geomagnetism.* Clarendon Press, Oxford.
Chapman, S. and Ferraro, V. C. A.	1940	*J. terr. Magn.* **45,** 245.
Christofilos, N. C.	1959	*J. geophys. Res.* **64,** 869.
Chubb, T. A., Friedman, H. and Kupperian, J.	1956a	*Scientific Uses of Earth Satellites* p. 147. Univ. of Michigan Press.
Chubb, T. A.	1956b	*Scientific Uses of Earth Satellites,* p. 152. Univ. of Michigan Press.
Chudakov, A. E.	1958	Report to IGY Symposium on Rockets and Satellites, Moscow.
Cornford, E. C.	1958	Report to IGY Symposium on Rockets and Satellites, Moscow.
Danilin, B. S., Mikhnevich, V. V., Repnev, A. I. and Shvidkovskii, E. G.	1957	*Usp. Fiz. Nauk.* **63,** 205.
Davis, R. J., Whipple, F. L. and Zirker, J. B.	1956	*Scientific Uses of Earth Satellites,* p. 1. Univ. of Michigan Press.
Drummeter, L. and Schach, M.	1957	Report to IGY Rocket and Satellite Symposium, Washington.
Dubin, M.	1956	*Scientific Uses of Earth Satellites,* p. 292. Univ. of Michigan Press.
Elliot, H.	1952	*Progress in Cosmic Ray Physics,* 455; ed. J. G. Wilson; North-Holland Publishing Co., Amsterdam.
Fejer, J. A.	1957	*Nature, Lond.,* **180,** 1413.
Freden, S. C. and White, R. S.	1959	*Phys. Rev. Letters,* 3, 9.
Friedman, H., Chubb, T. A., Kupperian, J. E. and Lindsay, J. C.	1958	IGY World Data Centre A. Report on Experimental results of U.S. Rocket Programme for IGY to 1958 July 1, 179.
Gringauz, K. I. and Zelikman, M. Ch.	1957	*Usp. Fiz. Nauk.* **63,** 239.
Groves, G. V.	1958	*Nature, Lond.,* **181,** 1055.
Groves, G. V.	1959	*Proc. Roy. Soc. A.*; **252,** 28.
Hagen, J. P.	1957	*Report to IGY Rocket and Satellite Symposium,* Washington.
Heppner, J. P.	1956	*Scientific Uses of Earth Satellites,* p. 234. Univ. of Michigan Press.
Hok, G., Sicinski, H. S. and Spencer, A. W.	1956	*Scientific Uses of Earth Satellites,* p. 263. Univ. of Michigan Press.
Imyanitov, I. M.	1957	*Usp. Fiz. Nauk.* **63,** 267.
Istomin, V. S.	1958	Report to IGY Symposium on Rockets and Satellites, Moscow.
Jacchia, L. G.	1958	Smithsonian Inst., *Astrophys. Obs. Spec. Rep.,* **19,** 1.
Jacchia, L. G.	1959	*Nature, Lond.,* **183,** 1162.
Jastrow, R. and Pearse, C. A.	1957	*J. geophys. Res.,* **62,** 413.
Jeffreys, H.	1952	*The Earth* (3rd Edn.), Cambridge University Press.
Jones, L. M. and Bartman, F. L.	1956	*Scientific Uses of Earth Satellites,* p. 85. University of Michigan Press.
Katz, L.	1956	*Scientific Uses of Earth Satellites,* p. 247. Univ. of Michigan Press.
King-Hele, D. G.	1958	*Proc. Roy. Soc. A.* **248,** 55.
King-Hele, D. G. and Leslie, D. C. M.	1958	*Nature, Lond.,* **181,** 1761.
King-Hele, D. G. and Merson, R. H.	1958	*J. Brit. Interplan. Soc.,* **16,** 446.
King-Hele, D. G.	1959a	Royal Society Discussion on Space Research, *Proc. Roy. Soc. A.,* **253,** 529.
King-Hele, D. G.	1959b	*Proceedings International Astronautical Congress, London.*
Kitchen, F. A.	1958	*Proc. Roy. Soc. A.* **248,** 63.
Kitchen, F. A. and Joy, W. R.	1958	*Nature, Lond.,* **181,** 1759.

KITCHEN, F. A., JOY, W. R. and BELLAM, E. R.	1958	Report to IGY Symposium on Rockets and Satellites, Moscow.
KOZAI, Y.	1959	Smithsonian Inst., *Astrophys. Obs. Spec. Rep.*, 22.
KRASSOVSKII, V. I.	1959	*Proc. I.R.E.*, **47**, 289.
KRAUS, J. D.	1958a	*Proc. I.R.E.*, **46**, 610.
KRAUS, J. D.	1958b	*Proc. I.R.E.*, **46**, 611.
KRAUS, J. D. and ALBUS, J. S.	1958	*Proc. I.R.E.*, **46**, 610.
KUPPERIAN, J. E., BYRAM, E. T., CHUBB, T. A. and FRIEDMAN, H.	1958a	IGY World Data Centre A, report on experimental results of U.S. Rocket Programme for IGY to 1958 July 1, p.190.
KUPPERIAN, J. E., BOGGES, A., MILLIGAN, J. C. and FRIEDMAN, H.	1958b	IGY World Data Centre A, Report on experimental results of U.S. Rocket Programme for IGY to 1958 July 1, p.194.
KUPPERIAN, J. E. and FRIEDMAN, H.	1958c	IGY World Data Centre A, Report on experimental results of U.S. Rocket Programme for IGY to 1958 July 1, p. 201.
KURNOSOVA, L. A.	1958	Report to IGY Symposium on Rockets and Satellites, Moscow.
LAGOW, H. E., HOROWITZ, R. and AINSWORTH, J.	1958	IGY World Data Centre A, report on experimental results of U.S. Rocket Programme for IGY to 1958 July 1, pp. 26, 38.
LAWDEN, D. F.	1955	*Astronautica Acta*, **1**, 185.
LESLIE, D. C. M.	1958	*Nature, Lond.*, **181**, 403.
LINES, A. W.	1958	*High Altitude and Satellite Rockets*, p. 89. Royal Aeronautical Society and British Interplanetary Society.
MARTYN, D. F.	1951	*Nature, Lond.*, **167**, 92.
MEREDITH, L. H., GOTTLIEB, M. B. and VAN ALLEN, J. A.	1955	*Phys. Rev.*, **97**, 201.
MITRA, S. K.	1947	*The Upper Atmosphere*, p. 150. The Royal Asiatic Soc., 1st edition.
MOULTON, F. R.	1931	*An Introduction to Celestial Mechanics*. New York, Macmillan Co.
MULLARD OBSERVATORY STAFF	1957	*Nature, Lond.*, **180**, 879.
NAZAROVA, I. N.	1958	Report to IGY Symposium on Rockets and Satellites, Moscow.
NEHER, H. V.	1952	*Progress in Cosmic Ray Physics*, 245.
NONWEILER, T.	1958	*Nature, Lond.*, **182**, 468.
O'KEEFE, J. A., HERTZ, H. G. and MARCHANT, M.	1958	Oblateness of the Earth by Artificial Satellites, Announcement Card 1408, Harvard College Observatory, June 24.
O'KEEFE, J. A., ECKELS, A. and SQUIRES, R. K.	1959	*Science*, **129**, 565.
OKHOTSIMSKII, D. E. and ENEEV, T. M.	1957	*Usp. Fiz. Nauk.*, **63**, 5; *J. Brit. Interplan. Soc.*, 1958 **16**, 263.
OKHOTSIMSKII, D. E. and ENEEV, T. M. TARATYNOVA, G. P.	1957	*Usp. Fiz. Nauk.*, **63**, 33.
PARKYN, D. G.	1958a	*Nature, Lond.*, **181**, 1156.
PARKYN, D. G.	1958b	*Nature, Lond.*, **182**, 787.
PLIMMER, R. N. A.	1958	Unpublished Ministry of Supply (U.K.) report.
PRIESTER, W.	1959	*Naturwiss.*, **46**, 197.
RATCLIFFE, J. A., SCHMERLING, E. R., SETTY, S. C. and THOMAS, J. O.	1956	*Phil. Trans.*, **248**, 621.
ROBERSON, R. E.	1957	*J. Franklin Inst.*, **264**, 181 and 269.
ROYAL AIRCRAFT ESTABLISHMENT STAFF	1957	*Nature, Lond.*, **180**, 937.
SADLER, D.	1958	*Proc. Roy. Soc. A.*, **248**, 45.
SCHILLING, G. F.	1958	Smithsonian Inst., *Astrophys. Obs. Spec. Rep.*, 14, 44.
SCOTT, J. M. C.	1957	*Nature, Lond.*, **180**, 1467.
SIRY, J. W.	1957	Report to IGY Rocket and Satellite Symposium, Washington.
SINGER, S. F.	1956	In *Vistas in Astronomy*, ed. A. Beer, Vol. 2, p. 878; Pergamon Press, London.
SINGER, S. F.	1959	*Proceedings Int. Astronautical Congress, London.*
SMART, W. M.	1953	Celestial Mechanics, Longmans, Green & Co., London.
SPITZER, L.	1950	*J. Brit. Interplan. Soc.*, **9**, 131.
SPITZER, L.	1956	*Scientific Uses of Earth Satellites*, p. 99. Univ. of Michigan Press.
STERNE, T. E.	1958	*Smithsonian Inst., Astrophys. Obs. Special Report*, **11**, 18.

Störmer, C.	1913	*Videnskap. Skrifter. Mat. naturw. Kl.*, No. 1–3.
Stroud, W. G., Hanel, R., Nordberg, W. and Stampfel, R.	1957	Report to IGY Rocket and Satellite Symposium, Washington.
Suomi, V. E.	1957	Report to IGY Rocket and Satellite Symposium, Washington.
Tabor, H.	1956	*Bull. Res. Council Israel, A: Math., Phys., Chem.*, 5A, **2**, 119.
Taratynova, G. P.	1957	*Usp. Fiz., Nauk.*, **63**, 51.
van Allen, J. A.	1955	*Phys. Rev.*, **99**, 609.
van Allen, J. A.	1956	*Scientific Uses of Earth Satellites*, p. 171. Univ. of Michigan Press.
van Allen, J. A. Ludwig, G. H., Ray, E. C. and McIlwain, C. F.	1958	*Jet Propulsion*, **28**, 588.
van Allen, J. A. and Frank, L. A.	1959a	*Nature, Lond.*, **183**, 430.
van Allen, J. A.	1959b	Rocket and Satellite Research Panel Report 2.
van Allen, J. A., McIlwain, C. E. and Ludwig, G. H.	1959c	*J. geophys. Res.* **64**, 877.
van de Hulst, H. C.	1947	*Astrophys. J.* **105**, 471.
Vernov, S. N. and Chudakov, A. E.	1959	Evening Lecture during IGY Symposium, Moscow. (Quoted in an article in *Pravda*, July 15).
Warwick, J. W.	1958	IGY World Data Centre A, Report No. 5.
Weekes, K. G.	1958	*Proc. Roy. Soc. A.*, **248**, 77.
Whipple, F. L. and Hynek, J. A.	1956	*Proc. I.R.E.*, **44**, 670.
Whitney, C. A.	1958	*Smithsonian Inst., Astrophys. Obs. Special Report*, **11**, 14.
Willmore, A. P.	1958	*Nature, Lond.*, **182**, 1008.
Yatsunskii, I. M.	1957	*Usp. Fiz. Nauk.*, **63**, 59.

CONTENTS

(Artificial Satellites)

		page
1.	Introduction	29
2.	Launching problems	32
3.	Perturbations of satellite orbits	38
4.	Methods for tracking satellites	57
5.	Temperature of a satellite and its control	62
6.	Meteorological studies with a satellite	65
7.	The measurement of atmospheric pressure and density	66
8.	Study of short-wave radiations from the Sun and other sources	68
9.	Cosmic and auroral radiations	71
10.	Magnetic measurements	74
11.	The ionosphere	78
12.	Study of micrometeorites	86
13.	The satellites already launched	88
	References	108

Dynamical Effects in the Motion of Earth Sputniks

L. I. Sedov
Academician, Department of Physics, University of Moscow

Investigation of the laws of motion of Sputniks is at present essentially carried out experimentally, using optical and radio observations. These provide the basis for theoretical computations of the orbit, as well as for the application of statistical methods adapted to the analysis of a large number of measurements.

In order to correlate the observed coordinates of orbital points of the Sputnik in space, careful tracking and extensive numerical data are required. Furthermore, accurate predictions of future positions have to be provided. The operation of the scientific equipment involved has to be planned in detail. A determination of the life-time of the various Sputniks and the solution of certain geophysical problems is also necessary, particularly those connected with the discovery and accurate measurement of anomalies in the gravitational field of the Earth, the density distribution in the Earth's atmosphere, etc.

Data on the rate of spin and orientation of the Sputnik in space are also of the greatest importance for many research problems, including the determination of the laws of interaction between the Sputnik and the atmosphere.

The theoretical computation of the orbit of a Sputnik and its motion around its centre of gravity requires consideration and analysis of the forces that act upon the Sputnik as a result of the Earth's gravitational field and the interaction of the Sputnik with the terrestrial atmosphere, as well as with the electromagnetic field.

From analyses that have already been carried out, it has been found that the perturbations of the motion of artificial satellites by the Sun and Moon, and by meteoric particles, are so very small that there is at present no need to take them into account.

The problem of the motion of a Sputnik, solely in terms of Newtonian gravitation, is a problem in celestial mechanics that has been extensively studied for a long time; it is in essence similar to the problem associated with the motion of the Moon and planets. If we choose a coordinate-system with origin at the centre of the Earth, which does not rotate with respect to the stars, then the orbit of the Sputnik during one revolution is very nearly the so-called "osculating ellipse".

The position of the orientation and plane of the osculating ellipse, with its focus at the centre of the Earth, can be described in the usual manner by the following parameters: The longitude of the ascending node Ω; the angle of inclination of the plane of the ellipse to the plane of the equator, that is, the inclination of the orbit i; and the distance between the perigee and the node, ω; (Fig. 1). The parameters that fix the size and the shape of the ellipse are its semi-major axis a and its eccentricity e. The laws of motion for one revolution are similar to Kepler's laws of motion for an osculating ellipse.

Because of the flattening of the Earth and of gravitational anomalies, the above-mentioned parameters are slowly modified by perturbations. The changes are characterized by secular variations and generally also by certain minor oscillations.

It is well known that the main secular effect of the Earth's flattening is to cause a uniform variation in Ω and ω. In the case of the first Soviet Sputniks the angle Ω changed by about 3° in 24 hr in the direction from East to West. During the same time the longitude of the perigee decreased by about $0^\circ.4$.

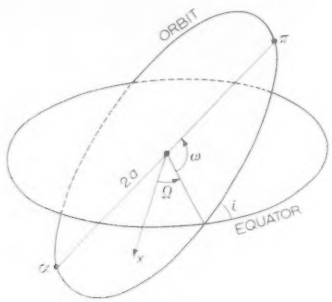


Fig. 1. Notations

Under the action of the Newtonian gravitational force some essential qualitative characteristics of the motion of a Sputnik as a solid body around its centre of gravity must vary greatly. They will do so in the same way which we are familiar with from the classical examples of the motion of the Moon or the Earth. A quantitative theory of Sputnik-motion will essentially have to take account of the diversity of the initial conditions and of the possible difference in the moments of inertia of the Sputnik about its central axes.

New significant effects are associated with the forces governing the interaction of the Sputnik with the atmosphere and with the electromagnetic field of the Earth.

The study of air resistance and its influence on the motion of the Sputnik has been dealt with in a number of papers.[1,2,3,4,5] Some conclusions as to the influence of these forces have been reached, but up to now the main difficulty has been the absence of sufficiently reliable data about the Earth's atmosphere at high altitudes.

The linear velocities resulting from the spinning of the Sputnik about its own centre of gravity are negligible in comparison with its velocity relative to the atmosphere. The influence of the spin needs therefore to be considered only because of the changing orientation of the Sputnik relative to the velocity vector of the Sputnik's centre of gravity.

In previously published papers the influence of aerodynamic forces has been examined on the assumption that the air-drag causes a force of resistance W, which is directed towards the velocity of the Sputnik's centre of gravity.

The determination of the resistance W involves the calculation of the coefficient C_x in the equation:

$$W = C_x \rho \, (S V^2/2), \qquad \ldots . (1)$$

where ρ is the density of the atmosphere at the particular point in space where the Sputnik is moving, S is the total area of its outer surface, and V its velocity.

The coefficient C_x is at present determined theoretically, through the application of the methods of molecular aerodynamics.

Generally speaking, this coefficient depends on the temperature of the atmosphere, the velocity of the Sputnik, and the mode of interaction of the molecules and atoms of the atmosphere with the particles belonging to the surface of the Sputnik. Moreover, C_x may also depend to a large extent on the geometrical form of the Sputnik and its orientation in relation to the velocity vector of its centre of gravity.

The theoretical analysis applied to the Soviet Sputniks, which took into account the spin around the centre of gravity, leads to numerical values of C_x which are either different from the true values, or vary with time up to 30 per cent (for Sputnik III).

If a charged Sputnik moves in the plasma, an additional resistance arises which, (in the case of considerable negative charges of the order of 30 V, small dimensions of the Sputnik, and a high altitude), might increase C_x by a factor of 2 (JASTROW and PEARSE 1957). It seems, however, that this effect is negligible for larger Sputniks.

The results also show that the resistance force due to interaction of the Sputnik with the magnetic field of the Earth is not significant.

Because of the small density of the atmosphere, ρ, the resistance force W is very small for altitudes exceeding about 230 km. According to recent data the density at such altitudes is of the order of $3 \cdot 10^{-13}$ g/cm^3.

For the Soviet Sputnik III the force of resistance at perigee is of the order of $W \sim 4$ grams and $W/G = 3 \cdot 10^{-6}$, where G is the weight of the Sputnik.

These published theories of the motion of the Sputnik which take resistance forces into account have been developed on the assumption that C_x is a constant, and that the density diminishes with altitude in accordance with some definite law. In particular, the exponential law of density variation is widely used, namely

$$\rho = \rho_0 \exp[-(Z - Z_0)/H], \qquad \ldots . (2)$$

where ρ_0 is the density at altitude Z_0 (Z being the altitude reckoned along the normal to the surface of the Earth's geoid). The quantity H may be regarded as a constant in many cases. However, the analysis of the observations, and the computation of the resistance forces for non-orientated Sputniks of non-spherical shape, indicate that the variations of C_x caused by the rotation of the Sputnik around its centre of gravity must be taken into account.

Analysis shows that the resistance force W causes secular changes of the orbit, which decrease the eccentricity and lower the perigee for an osculating ellipse, and cause, furthermore, a shortening of the period of revolution T.

The perigee is lowering slowly, the apogee more quickly.

The main drag-effect occurs in the near neighbourhood of the perigee, where the density is greatest, if the difference between the altitudes at the apogee and the perigee considerably exceeds H.

It should be noted that in this case the influence of aerodynamic factors on the rotation of the Sputnik is most prominent near the perigee. Let us suppose that the outer surface of the Sputnik is a surface of revolution, which is dynamically symmetric with respect to the axis of rotation. It is then known that the rotation of the Sputnik

about its axis undergoes a regular precessional motion, provided there are no perturbations relative to the centre of gravity.

It is evident that the aerodynamic forces cause some small perturbations. The resultant forces are always exerted on the axis of the Sputnik's rotation, and when this axis is shifted to its mirror-reflection relative to the plane Π passing through the velocity vector of the centre of gravity and the vector of the moment of momentum; they are represented by vectors which are mirror-reflections of each other in relation to this same plane Π.

It follows therefore that the vector of the moment of the aerodynamic force about the centre of gravity is perpendicular to the plane Π (when averaged over one period of the precessional movement). If the essential components of the aerodynamic moment act only in the region of the perigee, it follows from the above considerations that the secular term of the perturbed rotation should be a slow precession of the vector of the moment of momentum relative to a tangent to the orbit at the perigee.

This suggestion is also confirmed by detailed study of the perturbed motion.[7]

Analysis of the observations of Sputnik II has shown that the perturbation moments of the gravitational forces are comparable with the aerodynamic moments. The gravitational moments are known to produce a precession of the vector of the moment of momentum relative to the normal drawn to the plane of the orbit.

The combined action of aerodynamic and gravitational moments has been considered as well; the characteristic parameters of the perturbed motion have been derived, taking into account the regression of the orbit.[7]

The interaction of the Earth's magnetic field with the rotating metallic body of a Sputnik gives rise to Foucault currents in it, and thus to moments that act as a brake on the rotation of the Sputnik about the centre of mass; these moments must be taken into account by examining the rotation of a light non-stabilized Sputnik. In a number of cases the change of the laws of motion brought about by the magnetic field might influence the amount of resistance and therefore the life-time of the Sputnik, because of the variable orientation.

Since the moments of inertia about the different central axes of the carrier rocket of Sputnik I differ considerably from each other, and because of the accidental character of the initial perturbations, it is possible to assume that the angle of nutation will be near to 90°. The resulting motion resembles "somersaults". Indeed, observations confirm this conclusion. Analysis of the results obtained from investigations of the solar radiation permits certain conclusions to be drawn concerning the motion of Sputnik II (about the centre of mass) at the beginning of its existence, i.e., 1957 November 3. The parameters of the motion around this point are as follows:

Period of Precession	206 sec $\pm$ 10 sec
Velocity of Precession	1°.7 per sec
Angle of Nutation	86° $\pm$ 1°
Angle between the vector of the kinetic moment and the direction towards the Sun	90° $\pm$ 5°

We also possess observations of the brightness variations of Sputnik II, obtained at various stations in January 1958. The pronounced changes in brightness can be

fully explained if the general motion around the centre of mass is similar to the "regime of somersaults".

The period of the precession of the Sputnik will be equal to twice the period of the variations in its brightness, and indeed the optical observations agree very well with such a period. They show that the period of precession for the carrier rocket of Sputnik I is 80 sec, and for that of Sputnik III less than 15 sec.

The orientation of Sputnik III was derived from data obtained from the reduction of the readings of instruments which measure the variation in the position of the frame of the magnetometer with time. Preliminary results show that the type of motion of Sputnik III also proved to be close to one of pure somersaults. The period of precession was equal to 125–135 sec, while that of the actual rotation was considerably greater.

The results concerning the orbits of the Sputniks are based on the following data:

Sputnik I: 60,000 measurements from radio observations, and 400 from optical observations.

Sputnik II: 12,800 radio- and 2,000 optical-observations; 124 photographic plates are available.

Sputnik III: (up to 1958 July 7): 52,750 radio measurements and 1,260 optical observations.

These data were analysed by high-speed electronic calculating machines which led to the orbital elements.

Table 1. The parameters of the initial orbits.

	Sputnik I	Sputnik II	Sputnik III
Period of revolution around the Earth, T	96·17 min.	103·75 min.	105·95 min.
Minimum height (Perigee distance), h_{min}	226–228 km	225 km	226 km
Maximum height (Apogee distance), h_{max}	947 km	1671 km	1881 km
Inclination of orbit, i	65°·129	65°·310	65°·188
Rate of regression of line of nodes (degrees per day), $\Delta\Omega$	−3°·157	−2°·663	−2°·528
Rate of rotation of perigee (degrees/day), $\Delta\omega$	−0°·432	−0°·407	−0°·326
Rate of decrease of period (seconds per day), ΔT	1·8 sec/day	3·08 sec/day	0·75 sec/day

Table 1 lists the parameters of the orbits at the beginning of the motion. Except for small regions near the Earth's poles, the orbits of the Sputniks traverse all parts of the upper atmosphere. Thus great opportunities for research of these regions are opened up.

Fig. 2 represents the variation of the periods.

Fig. 3 demonstrates the variation of the minimum altitude h_{min}, and Fig. 4 shows the variation of the maximum altitude of the trajectory, h_{max}.

The calculation of the orbit and the consequent predictions are rather complicated and tedious, since they necessarily involve the analysis of a very large number of observational data, and demand of course great precision.

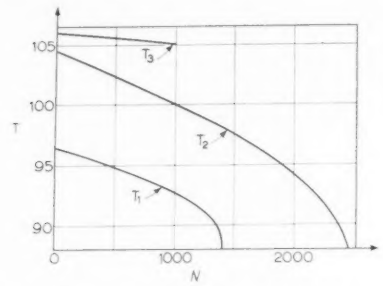


Fig. 2. The change of the orbital periods of Sputniks I, II and III; the periods T are plotted against the number of orbital revolutions N.

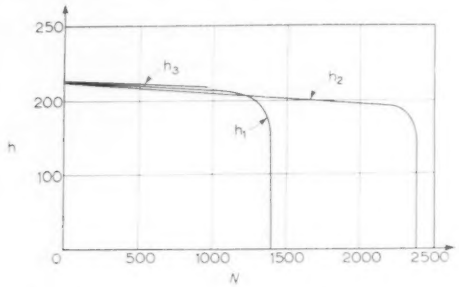


Fig. 3. The heights of the Sputniks at perigee plotted against the number of orbital revolutions.

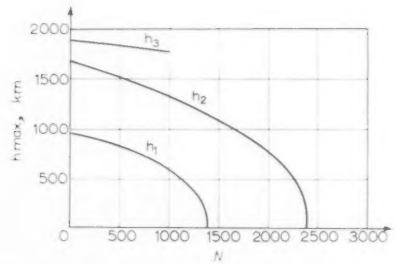


Fig. 4. The heights of the Sputniks at apogee ($h_{\max}$) plotted against the number of orbital revolutions.

Predictions are required as soon as possible after receipt of the observational data. Special methods have been developed that permit the derivation of these results at the required speed, so that we now have successful methods of accurate operational prediction at our disposal.

The results of the observations of the revolution of the Sputnik's orbit can be used to determine the density of the Earth's atmosphere.

From the equation of the motion of the centre of gravity of the Sputnik, the following formula for the change of the revolution-period ΔT during one revolution can be derived:

$$\Delta T = -\frac{3a^2\pi S}{\sqrt{(m\mu)}} C_x \rho_\pi \sqrt{\frac{2\pi H}{e}}\, f[e, (H/ae)] \qquad \ldots . (3)$$

where $\mu = 3{\cdot}98600 \times 10^{20}$ cm^3/sec^2, m = mass of the Sputnik, e = eccentricity, a = semi-major axis of the osculating ellipse, and $f(e, H/ae)$ is a known function.

The eccentricity e at any moment can be determined by the parameters of the orbit; ρ_π is the density at perigee.

For small values of e and for $(H/ae) < \frac{1}{4}$ the asymptotic value of the function f is equal to 1. The parameters $C_x\rho_\pi\sqrt{H}$ enter as factors in the right-hand part of equation (3).

The value of ΔT, of the eccentricity e and of the semi-major axis a are determined directly from the observations; the value of H may be evaluated from the temperature of the atmosphere in the region of the perigee. From this and the theoretical data about the coefficient C_x, the density ρ_π can be determined. The accuracy of such a computation depends on the reliability of the values adopted for C_x and H.

The determination of H from orbital data is also possible; but to determine it we need to know the very small value $\Delta h_{\min}$ with great accuracy; and thus this calculation does not lead to a reliable result.

If $H = 25$ km and $C_x = 0{\cdot}525$, the following range of values are obtained for the density: $h_{\min} = 225$ km, $\rho_\pi = 3{\cdot}10^{-13}$ to $4{\cdot}10^{-13}$ g/cm^3; (see References 8, 9).

These values should be regarded as mean values of the air density at the latitude of the perigee at a great number of points uniformly distributed along the parallel.

The calculations which have been carried out so far and which take into account all the data obtained for the evolution of the orbits of all Sputniks, show a systematical increase of the product $C_x\rho_\pi\sqrt{H}$ for fixed altitudes with the transition of the perigee from night to day, due to the precession of the orbit.

A hypothesis has been suggested that this is caused by the increase of the density of the atmosphere at an altitude of 225 km with the transition from night to day.

The above values for the density are in good agreement with data obtained from rocket measurements.

REFERENCES

1. Okhotsimskii, D. E., Eneev, T. M., Taratynova, G. P. Determination of the maximum existence of an artificial Earth satellite, and researches on the secular perturbations of its orbit. *Usp. Fiz. Nauk.*, **63,** 1a, 33–51, September 1957.

2. Taratynova, G. P. On the motion of an artificial satellite in the nearly central field of gravitation of the Earth, and in the presence of appreciable atmospheric drag. *Usp. Fiz. Nauk.*, **63,** 1a, 51–59, September 1957.

3. Yatsunskii, I. M. On the influence of geophysical factors on the motion of an artificial satellite. *Usp. Fiz. Nauk.*, **63,** 1a, 59–73, September 1957.

4. Krause, H. C. L. Die säkularen und periodischen Störungen der Bahn eines künstlichen Erdsateillten. *Proceedings of the VIIth International Astronautical Congress*, pp. 523–587, Rome, 1956.

5. de Nike, J. The effect of the Earth's oblateness and atmosphere on a satellite orbit. *Proceedings of the VIIth International Astronautical Congress*, pp. 507–22, Rome, 1956.

6. Jastrow, R., Pearse, C. A. Atmospheric drag on the satellite. *J. Geophys. Res.*, **62,** 3, 1957.

7. Beletskii, V. V. Researches on the motion of an artificial Earth satellite about its centre of mass. *Artificial Earth Satellite*, No. 1, Moscow, 1958.

8. Elyasberg, P. E. Limits of the upper atmosphere from the data of an artificial satellite orbit. *Artificial Earth Satellite*, No. 1, Moscow, 1958.

9. Lidov, M. L. Results concerning the atmosphere from observations of the motion of the first artificial Earth satellite. *Artificial Earth Satellite*, No. 1, Moscow, 1958.

Photographic Observations of Artificial Earth Satellites

A. G. Masevich and A. M. Lozinskii

The Astanomical Council of the USSR Academy of Science, Moscow

1. Introduction

The purpose of photographic observations of artificial Earth satellites (supplying a photograph of the track of a satellite and of the stars in its neighbourhood) is to mark with the maximum precision the moment of time corresponding to the position of the satellite at a certain point of its track.

Photographic observations of artificial Earth satellites are considerably more accurate in their coordinate determinations than the radio observations. Photokine-theodolite observations are very valuable, but maximum precision can only be achieved when a photograph of the satellite is taken against the background of stars.

As cosmic bodies, artificial Earth satellites have raised new problems in the field of photographic astrometry.

Unlike observations of stars, and even minor planets and comets (for which it is sufficient to know the time of observation approximately), satellite observations require (because of the great apparent angular velocity of the satellites) the moment of exposure to be fixed to within $0^s{\cdot}001$. In fact the precision of satellite observations is determined rather by the precision of the time-determination, than by the precision of coordinate-measurements. Therefore special consideration should be given to the registration of the time.

No astrographs in current use have facilities for measuring time to an accuracy of greater than one second. For the solution of the problem of photographic satellite observations it is necessary to create new systems of shutters for astronomical telescopes. Because of the relatively fast motion of the satellite, the effective time of exposure is a fraction of a second. In order to increase this time, the telescope or the plate should be moving with the satellite. This problem has been solved in the Baker-Nunn cameras in the U.S.A. (Henize, 1957).

Another peculiarity of artificial satellites (contrary to those of usual celestial bodies) is the relatively rapid alteration of their orbital elements in consequence of the atmospheric drag and the deviation of the gravitational field of the Earth from the spherical field. Recent investigations show the atmospheric drag to be a complex function of many parameters, the importance of which should be found out with the help of very precise satellite observations. Highly accurate observations are necessary also for the solution of geodetic problems. For the study of orbital evolution many observations of similar type from different places on the Earth's surface are needed. This requires the creation of a special satellite service.

2. Stations for Photographic Observations, and their Equipment

When photographs of stars (point-sources) are being taken, the effective power of the telescope depends on the area of the objective lens and on the image-spread,

but does not depend on the focal-ratio of the objective lens. The effective power is proportional to the ratio D^2/d^2, where D is the diameter of the objective lens and d the diameter of the image. When photographs of a satellite are being taken, the effective power of a telescope is determined by the ratio D^2/dF, as the satellite image moves along the plate with a velocity which is inversely proportional to the focal ratio of the camera.

After the launching of the first Soviet satellite, the Astronomical Council of the USSR Academy of Sciences started the organization of special stations for photographic observations, in addition to the net of visual stations which had been working already (Masevich and Lozinskii, 1958). It was decided to use existing aerial photo-cameras.

Of all the existing aerial photo-cameras the camera "NAFA 3 c/25" turned out to be the most suitable for photographing Earth satellites (brighter than the 3rd magnitude).

The camera, supplied with an objective lens of the type "Uran–9" ($F = 25$ cm, $D = 10$ cm, field of view $30^\circ \times 50^\circ$), was provided with a special shutter.

To obtain sharp end-points of the satellite track, the times of the opening and closing of the shutter were decreased as much as possible and brought to 0·002–0·003 sec by strengthening the springs.

A specially developed electric device allows the opening and closing of the shutter to be carried out by hand, with the help of a button, in two phases: the first phase consists of the opening of the shutter at the start of the impulse, and its closing at the end; the second phase comprises the opening at the start of the impulse and the closing when the impulse is repeated.

The kinematics of the device was changed in such a way that the time necessary for the raising of the shutter did not exceed $0^s{\cdot}5$. This allows the opening and the closing of the shutter to be carried out with the help of a contact chronometer. In this case a current, not exceeding 2 mA, runs through the contacts of the chronometer.

A special contact, together with the chronograph, is used to record the moments of opening and closing of the shutter. A coincidence (with the precision of $0^s{\cdot}002$) of the midpoint of the exposure with the arithmetic mean of the moments of opening and closing of the shutter can be reached by regulation of this contact, by means of a device supplied with a photocell. This excludes the influence of the delay between the actual moment of contact and the times of the complete opening and closing of the shutter. Such a recording of the time has the advantage that it avoids a number of systematic errors due to measuring the edges of a satellite track on the negative.

To obtain several exposures on one frame, automatic transport of the film after every exposure was eliminated. Furthermore, the camera is supplied with a frame view-finder and a special support, which makes it possible to turn the camera about a horizontal axis into the necessary position and to fasten it afterwards. If several short exposures (about $0^s{\cdot}5$) are made during the time of the satellite's passage through the field of view of the camera, the image of the satellite will consist of several tracks of 2 mm length, while the star images will be of the form of extremely short trails practically like points.

Knowing the positions of the end-points of the satellite-tracks with respect to the stars, it is possible to determine the position of the satellite on the celestial sphere at certain periods of time.

Twenty-six stations belonging to astronomical observatories and physical and mathematical departments of several universities were equipped with such cameras. A list of the photographic stations of the USSR and their coordinates are given in Table 1 on page 125.

As experience shows, these cameras permit the determination of the position of bright satellites with a precision of $0{\cdot}1'$ of arc, as well as a corresponding accuracy for the time-coordinate.

At those stations which have no time-service, a recording chronograph with a quartz generator, which receives time-signals before and after the observation of the artificial Earth satellite, is used as a time-keeping device. In this case the error involved in the comparison of the recording chronograph with the precise time-signals should not exceed 0·005 sec. For this purpose an additional device on the receiver is used, which allows the marking of the time-signals to be made directly onto the tape of the chronograph. The delay of a signal in the receiver and in the additional device is accurately measured. It is also possible to obtain direct measurements of the delay of the signals in the additional device during the time when radio signals are received.

Photographs are taken with a fixed camera. Experience shows that a satellite's track, consisting of several parts, obtained by taking a number of short exposures ($0^s{\cdot}2$–$0^s{\cdot}1$), is most favourable. Such a photograph also shows star images in the form of very short trails.

3. Precise Reduction

The track of a satellite on the plate consists of a number of point images lying practically on a line. On both sides of any point-image two reference stars S_1 and S_2 (with α and δ known from the catalogue) may be chosen (Fig. 1). It is possible

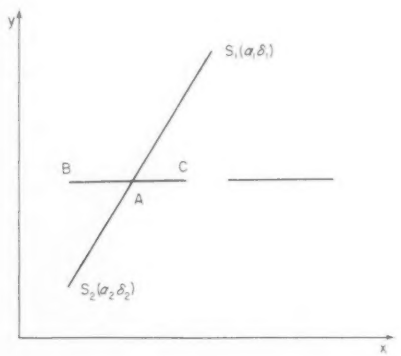


Fig. 1. Satellite track with two reference stars for precise reduction.

further to orientate the photograph in the measuring machine in such a way that one of the axes (e.g., x) can be directed along the track. In this case the rectangular coordinates of the points S_1 and S_2—the coordinate y of the point A of intersection of the straight line S_1S_2 with the track and the coordinates x of the ends of the track B and C—can be measured. Using the fact that the three points A, S_1 and S_2 are on the same straight line, the coordinate x of the point A can be calculated.

The time at which the satellite passed the point A is obtained from the known times at which the satellite passed the points B and C, and from the ratio of the segments BA and AC.

One may determine the equatorial coordinates of a satellite quite simply by a method proposed by A. N. Deutsch (1958) (for $\delta \leqslant 60°$, when the distance between the reference stars is $\leqslant 5°$ and the distance from the optical centre $\leqslant 15°$). In several cases third-order terms have to be taken into account, especially for large declinations. A method proposed by A. A. Kiselev (1959) is somewhat more complicated, as far as calculations are concerned, but suitable for any declination. Both methods give the position of the satellite from the equatorial coordinates of the reference stars directly, without any preliminary calculations of the ideal coordinates.

Altogether 265 photographs of the second Sputnik, 1957β, (144 of the photographs give maximum precision), 1529 photographs of the rocket-carrier of the third Sputnik 1958δ, (874 of the plates give highly precise results) and 185 photographs of the third Sputnik itself 1958δ_2 (56 of these are precise) have been obtained with the NAFA-cameras.

The comparatively small number of photographic observations of the third Sputnik, 1958δ_2, is due to its faintness (it was usually fainter than 5 mag.), while the NAFA-cameras at the observing stations were initially designed for photographing only bright objects (about 3–4 mag.).

4. Photography of Faint Satellites

Many observatories in the Soviet Union are equipped with telescopes suitable for photographing satellites fainter than 3 mag.; for example the astrographs of the Crimean Astrophysical Observatory of the USSR Academy of Sciences and at the Southern Station of the Sternberg Astronomical Institute, which both have an aperture of 40 cm, with $F = 160$ cm and rather large fields of view, i.e., $10° \times 10°$.

There are some considerably more powerful cameras at the Alma-Ata, Abastumany, Burakan, and at other observatories. Unfortunately, they cannot be used

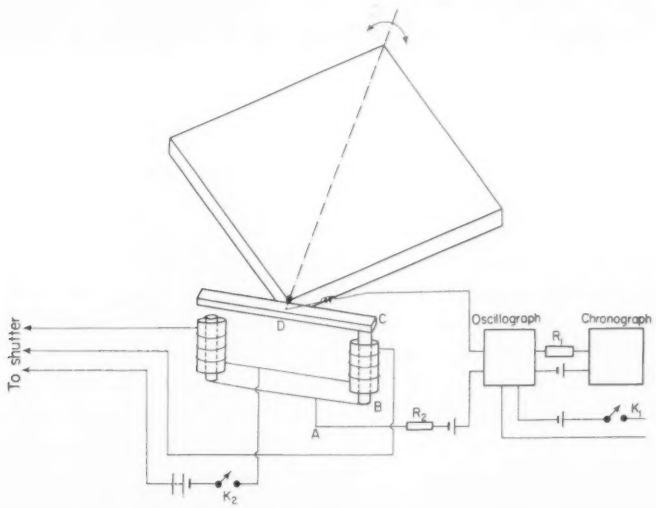


Fig. 2. Timing-device for the Alma-Ata Maksutov Telescope.

immediately for precise photographic observations, because none of them has shutters with a velocity of opening and closing rapid enough for precise time markings. As a rule, celestial objects are photographed with these telescopes with exposure times of one hour or so, and therefore the shutters are usually opened by hand. Besides that, it is necessary to take into account the fact that there are no quick-working shutters for large apertures, and that their construction is very complicated.

Nevertheless, a number of observatories have tried to apply the existing telescopes to satellite observations.

The following device has been made for the Maksutov meniscus telescope in Alma-Ata (F = 1200 mm; D/F = 1 : 2·4) (Rozhkovskii, Matyagin and Charitonov). In front of the corrector of the telescope there is a plane-parallel glass plate, to which an oscillatory motion around an axis parallel to the visible satellite motion is

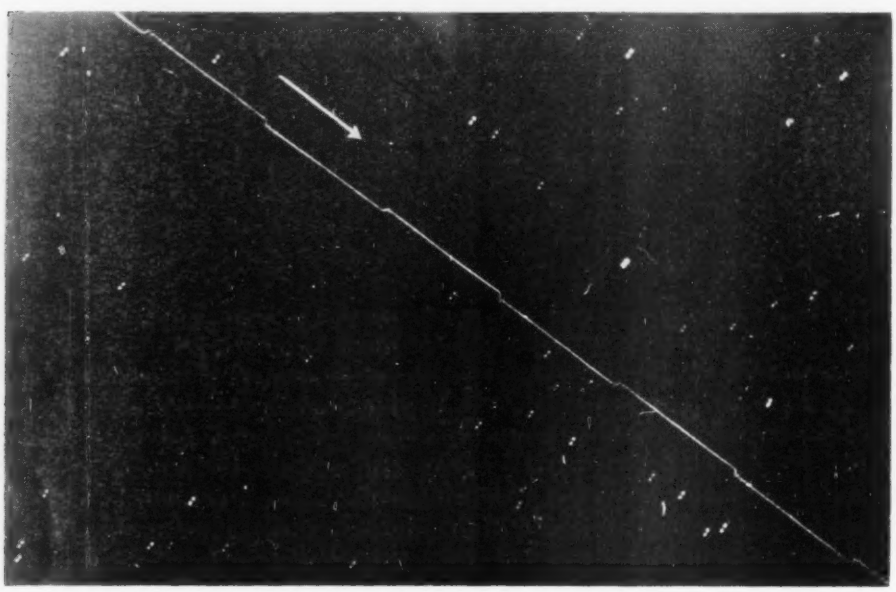


Fig. 3. Photograph of a satellite obtained with the Maksutov Telescope of the Alma-Ata Astrophysical Observatory.

given (Fig. 2). The transition time of the plate from one position to another does not exceed 0·01–0·02 sec, these times being recorded by an oscillograph, which simultaneously registers the time-signals. Due to the oscillation of the plate, the track of a satellite appears as a broken line (Fig. 3). The position of the breaks determines the satellite position relative to the stars and the corresponding moments of time.

A special plate-holder for astronomical telescopes, allowing precise timing of a satellite's passage through certain points of the plate, has been constructed at the Engelhardt Astronomical Observatory (near Kazan). A plane-parallel glass plate with a network of parallel lines is placed in front of the photographic plate. During the exposure, the frame, together with the glass-plate, is shifted relative to the photographic plate and the position of the network is marked at certain intervals of time on the photographic plate by printing their shadows with the aid of a flash-light. The velocity of the shift is chosen in such a way that 20–30 mm are traversed

during the time of exposure. The track of a satellite on the photographic plate is interrupted in those places where its image is cut by the network. The exact time of these interruptions is determined by measuring their positions and the positions of the network at fixed moments of the flashes. A similar method of photography has been worked out at the Sternberg Astronomical Institute in Moscow.

L. A. Panayotov (1958) at Pulkovo has constructed a special camera for photographic observations of faint satellites: the film moves in this camera with the same speed and in the same direction as the image of the satellite in the focal plane. There are two rectangular apertures of different width in the frame of the camera, and narrow slits (on different levels) in front of the right edges of the frame and in front of each aperture above and below. Through these slits the film is exposed with a neon lamp which flashes each second owing to the closing of the contacts of the chronometer.

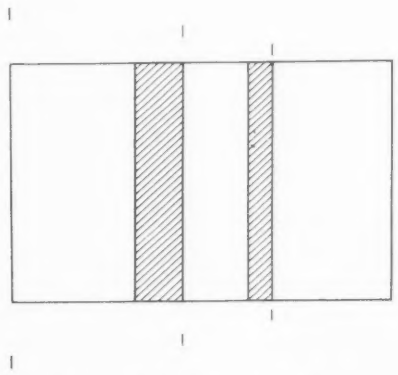


Fig. 4. Focal plane of the camera with the moving film.

The field of background stars against which the frame and the screens are seen is photographed on the fixed film before (and after) the passage of the satellite. Then the film is set moving and the photographs of the satellite with the corresponding timing are taken. If the satellite moves quicker (or slower) than the film, its image consists of three spots, instead of one. By measuring the positions of the beginnings of these spots, referring to certain time-markings, we obtain three moments of time. Their position referred to the frame and the stellar background enables their coordinates in the sky to be computed.

It is possible to obtain photographs of satellites up to the 6th or 7th stellar magnitude with such cameras.

An interesting method for photographing faint artificial Earth satellites has been proposed by Abele (1960) of Riga University. By means of a special plate-holder (Fig. 5) it is possible to shift the photographic plate uniformly for some millimetres in the direction of motion of the satellite image with a velocity which is equal to that of the image in its passage across the focal plane of the camera. During this time the image of the satellite affects practically the same part of the photographic emulsion. The plate remains motionless for some time, and then quickly returns to its initial position. Then again it remains motionless. After that the cycle is repeated. When the plate-holder is motionless, the satellite image passes quickly across the plate; and if the satellite is faint its passage does not register on the emulsion. During this

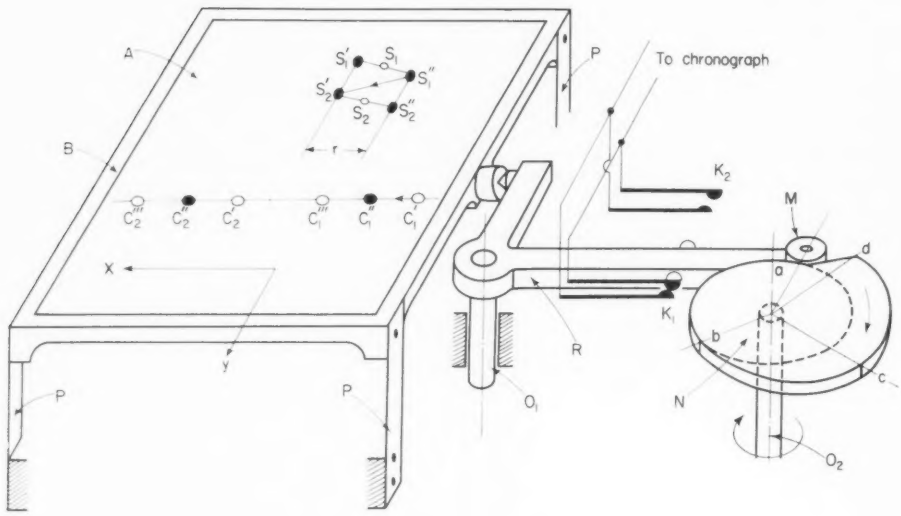


FIG. 5. General scheme of the plateholder designed by ABELE.

time the faint background-stars, the images of which were shifted while the plate-holder was moving, are exposed.

To move the plate-holder, a cam of a special form is used. The cam is set into rotation by a synchronous motor, which is fed with alternating current from a noise generator. The same cam sends signals to the chronograph at the moments when the plate-holder changes its motion. This results in the formation of 10–15 separate images of the satellite on the plate. The image of each star has the form of two parallel

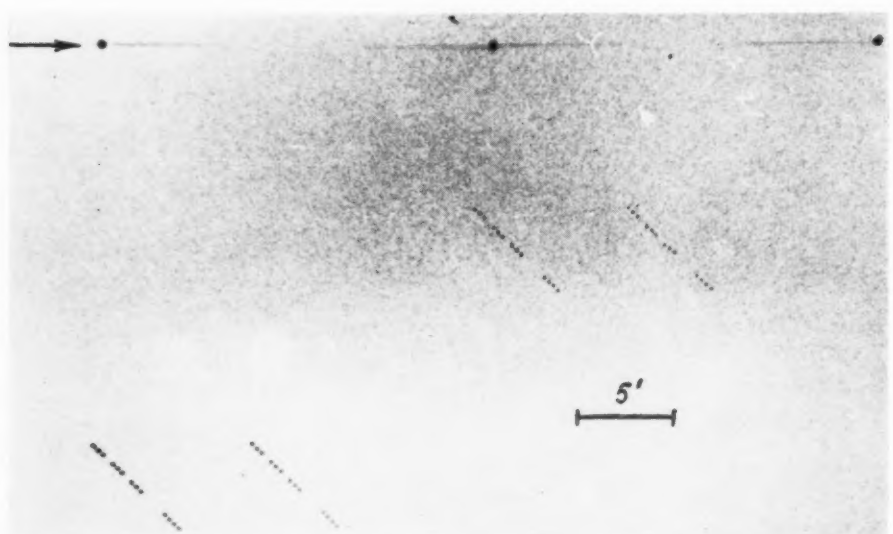


FIG. 6. Photograph of a satellite obtained with a camera carrying the plateholder designed by ABELE.

Table 1

List of the USSR stations for photographic observations of artificial Earth satellites

Code Number	Location of the Station	Institutions responsible for Observations	Coordinates of the Station		
			Latitude	Longitude	Altitude (m)
1	2	3	4	5	6
003	Abastumany	Abastumany Astrophysical Observatory	41° 45′ 18″	42° 49′ 30″	1600
007	Baku	State University, Azerbaijan	40° 22′	49° 49′ 30″	84
079	Irkutsk	Irkutsk State University	52° 15′ 44″·1	104° 20′ 43″·5	468
023	Kiev	Kiev State University	50° 27′ 11″·43	30° 30′ 07″·05	184
028	Crimea	Astrophysical Observatory of the USSR Academy of Sciences	44° 43′·7	34° 01′ 30″	569
030	Leningrad	Leningrad State University	59° 56′ 38″·4	30° 16′ 15″	15
031	Lvov	Lvov State University	49° 49′ 57″·6	23° 46′ 46″·95	330
034	Moscow	Sternberg Astronomical Institute	55° 41′ 58″	37° 32′ 40″·5	200
039	Pulkovo	The Pulkovo Observatory of the USSR Academy of Sciences	59° 46′ 13″·6	30° 19′ 42″·15	76
041	Rostov-Don	Rostov State University	47° 13′ 31″	39° 42′ 40″	68
045	Sverdlovsk	The Urals State University	56° 49′ 25″·8	60° 36′ 28″·8	274
054	Tomsk	Tomsk State University	56° 28′ 06″·3	84° 56′ 48″	114
055	Ozhorod	Uzhorod State University	48° 38′ 03″	22° 18′ 01″·2	
060	Kharkov	Kharkov State University	50° 0′ 10″·9	36° 13′ 57″	138
067	Alma-Ata	The Astrophysical Institute of the Kazakh Academy of Sciences	43° 11′ 16″·62	76° 57′ 27″·6	1450
068	Stalinabad	The Astrophysical Institute of the Tajik Academy of Sciences	38° 33′ 39″·9	68° 46′ 52″·08	820
070	Kiev	Main Astronomical Observatory of the Ukrainian Academy of Sciences	50° 21′ 55″·84	30° 29′ 52″·5	184
072	Moscow, Astrosovjet	Astronomical Council of the USSR Academy of Sciences	55° 42′	36° 47′	
073	Odessa	Odessa State University	46° 28′ 37″	30° 45′ 30″	53
074	Ashkhabad	Institute for Physics and Geophysics of the Turkmen Academy of Sciences	37° 57′ 18″	58° 20′ 45″	234
075	Tashkent	Astronomical Observatory of the Uzbek Academy of Sciences	41° 19′ 30″	69° 17′ 37″·05	479
076	Kazan, Station "Observatory"	Engelhardt Astrophysical Observatory	55° 50′ 26″·2	48° 48′ 56″·1	98
077	Nikolajev	Branch of the Main Astronomical Observatory, USSR Academy of Sciences	46° 58′ 18″·0	31° 58′ 26″·25	518
082	Burakan	Burakan Astrophysical Observatory	40° 19′ 56″	44° 15′ 55″	1480
083	Tartu	Astronomical Department of the Institute for Physics and Astronomy of the Estonian Academy of Sciences	58° 22′ 47″·16	26° 43′ 17″·7	67
084	Riga	Astronomical Department of the Physical Institute of the Latvian Academy of Sciences	56° 57′ 08″·03	24° 07′ 01″·26	39
1	2	3	4	5	6

lines of spots, and the satellite's positions appear as separate spots (Fig. 6). The time corresponding to the position of the satellite on the plate is obtained from the times of the beginning and the end of the uniform shift of the plate-holder in the direction of motion of the satellite-image.

The position of each star at this time is equal to the mean position between two corresponding spots from two lines of its images, if the shape of the cam is such that $t_b - t_a = t_d - t_c$; (Fig. 5).

If the motions of the satellite and the plate movement are well synchronized, it is possible to obtain a gain in the limiting magnitude for a satellite of up to 3–4 mag.

REFERENCES

Abele, M. K.	1960	*Astr. Zh.*, **37**, No. 1.
Deutsch, A. N.	1958	*Astr. Zh.*, **35**, No. 5.
Henize, K. G.	1957	*Sky and Telescope*, **16**, No. 3.
Kiselev, A. A.	1959	*Astr. Zh.*, **36**, No. 2.
Masevich, A. G. and Lozinskii, A. M.	1958	*Publ. Astron. Soc. Pacific*, **70**, 412.
Panayctov, L. A.	1958	Russian Astronomical Circular, No. 199.
Rozhkovskii, D. A., Matyagin, V. C. and Charitonov, A. V.	1958	*Astr. Zh.*, **35**, No. 3.

Archaeomagnetism

An Indirect Study of Geomagnetic Secular Variation

J. C. Belshé

Department of Geodesy and Geophysics, University of Cambridge.

Summary

In recent years techniques have developed which make possible the accurate sampling and measurement of the weak permanent magnetism in Earth materials. Utilizing the magnetism acquired by ancient kilns in the course of their operation, a record of the secular variation of the geomagnetic field in Europe is being obtained. Most of the work reported concerns England during the 1st to 16th centuries A.D. A record extending back 5000 years may eventually be attained. The method also holds promise to archaeologists of a new way to date sites. Other indirect methods are discussed for extending the secular variation knowledge back far beyond the period during which it has been directly observed.

1. Secular Changes of the Earth's Magnetic Field

The Earth's magnetic field approximates to that of a uniformly magnetized sphere centred within the Earth. Irregular features are exhibited as well which may cause large departures from the regular field in some localities. Moreover, at any place on the Earth the field is constantly changing.

Modern knowledge of the geomagnetic field recognizes that it exhibits an extremely wide range of variation periods, with respect to time. At the one extreme are changes which take place on a time scale of seconds. At the other extreme, studies of the magnetization acquired by rocks in remote geological times suggest changes which would only be apparent after times of the order of several million years. The short period changes reflect disturbances of the ionosphere or of solar origin. The long periods are associated with changes in the Earth's core, which is the source of the Earth's dipole field, and the orientation of the magnetic dipole with respect to features on the Earth's surface.

Between these extremes many periods are recognized. One has a time scale of the order of 10^3 years and is termed the secular variation. It is this change, the documentation of its details, period, and amplitude, which concerns us in this paper.

Within 60 years of the first measurements of the geomagnetic field's directions at London, the change of declination with time was recognized. This secular change of direction of the magnetic force can be easily described by a graph constructed for a particular place of observation. Such is Fig. 1 which shows the change in declination and inclination observed at London since Borough's first measurements. A few European observatories have recorded these data for four centuries and several elsewhere have extended this knowledge during the last two centuries. Even so, the study is very incomplete. If Fig. 1 is interpreted as showing 3/4 of a secular variation cycle, then a period of about 530 years is suggested, but the records from other

observatories would suggest other periods. Speaking of the period as approximately 10^3 years expresses correctly the order of magnitude of the time of variation and is in accordance with findings described in this paper.

Two features emerge from a study of the secular variation record. First, the phenomenon is a regional one and extends over areas of the Earth's surface of continental dimensions, 2000 to 5000 kilometres in diameter. Within such a region the rate of change of the magnetic field forms a consistent pattern, and over a few years the change of a magnetic component at a particular place can be adequately described by a linear function. Second, the regional centres of change drift westwards slowly at a rate of approximately one degree in five years.

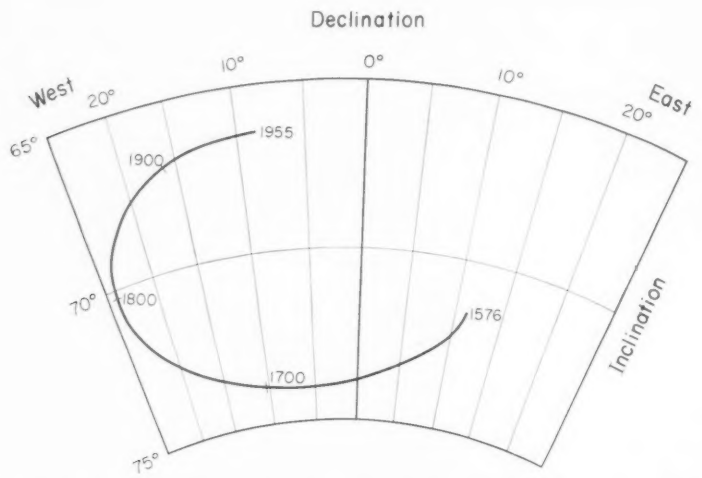


FIG. 1. The secular change in the magnetic force direction observed at London (after Bauer).

The dynamo theory of the Earth's magnetic field is now the most favoured explanation of this phenomenon. This theory assumes convective motion in the fluid, metallic (and thus electrically conductive) core of the Earth. Patterns of motion of the core material have been described which could result in a dynamo of rotational symmetry which would give the geomagnetic dipole field we observe at the Earth's surface. In this framework the regional secular disturbances of the magnetic field are considered as originating in the outermost layers of the Earth's core, possibly as eddies of the fluid motion.

For a discussion of the secular variation the reader is referred to CHAPMAN and BARTELS (1940), particularly to Chapters 3, 18 and 26; and to VESTINE, *et al.* (1949 and 1950). The last of these contains charts which well document the variation by epochs from 1905 to 1945. ELSASSER (1955 and 1956) has surveyed the dynamo theory, and BULLARD, *et al.* (1950) analysed the westward drift.

2. ARCHAEOMAGNETISM

(*a*) *Theory*

The firing of clay in making pottery or bricks is an activity of man which results in permanently magnetized material. This is because most natural clays contain

appreciable amounts of iron oxides, as witnessed by the red or brownish colour of bricks and tiles. These oxides are naturally ferromagnetic with a Curie temperature range of 500°–700° C. To fire the clays into ceramic objects requires that they be heated to temperatures in excess of 850° C. Thus the firing process destroys any earlier magnetization the clay possessed. The new magnetization will have a direction and intensity related to the magnetic field in which occurred the cooling through the Curie temperature, and this will, in nearly all instances, be the geomagnetic field.

Observation shows that this magnetism is stable and uniform both in direction and intensity. These characteristics are to a great extent due to the fact that the magnetic oxides are well scattered throughout the clay as very small particles, and that the slight vitrification of the clays which takes place during the firing inhibits later chemical changes.

In this manner an indirect record of the magnetic field from past times can be preserved. This will, of course, be a record applicable only to a particular place and time, and, unless one can be sure of the orientation of the sample as it cooled the intensity of the remanent vector is the only magnetic component which can be ascertained. For these reasons the investigations of tiles, bricks and pottery would not seem likely to be rewarding. But the structures inside which these objects were fired satisfy very well this requirement. Such structures are the ceramicists' kilns and, until fairly modern times (and still today in less advanced societies) these were made from a pit dug in the ground. These kilns might have quite varied modifications for their internal structure, their lining, the flue opening, etc., but they all made use of the earth-pit foundation for structural support. The walls and internal structures of the kilns have been fired just as well as the shaped clay objects they contained. Being an integral part of the ground, we can be sure of the structures' orientation, and observation warns us of an area where sagging or collapse may have occurred. The method of construction also favours preservation of at least the foundations of the structure.

The magnetic field at the time of the last firing should be the one recorded. The difficulty of encountering quite different magnetizations within one structure is minimized by the observation that these kilns were normally used for only short periods, perhaps as few as three firings on the average. This fact is due to the destructive expansion and contraction of the weak clay structure during firing and to the simplicity of the construction which favoured making a new kiln rather than extensively repairing a used one.

Enough kilns are found in each archaeological field season to give promise of an extensive record being pieced together within a few years. On an average, about 25 Roman kilns are found in Britain each year. During the few decades that a particular cultural centre was active a subtly varying succession of ceramic products was evolved. Such remains are often found in association with the kilns excavated. They are the basis from which the archaeologist works out the age succession of a kiln series and this, combined with historic record or absolute dating techniques, provides the chronological order of the whole record.

Man has been making fired pottery extensively since at least 3500 B.C. It may be possible to find, in the remains of this work, a record of the secular variation extending back 5000 years, some five times the time period of the variations and twelve times the length of the observational record.

(b) Techniques

Several research workers, beginning in the mid-nineteenth century, have attempted to use pottery for a study of this nature. It was E. THELLIER (1938) who first recognized the potential advance possible by concentrating on kilns and who developed, and described, suitable techniques for the research. The techniques described here follow many of the practices of that worker who has generously aided and advised the author and his archaeological collaborator Mr. R. M. COOK, in their study.

FIG. 2. Preparing oriented samples:
Plaster caps having just been poured over samples from a late Hellenic kiln in Keramicro, Athens.

The remanent magnetism must be measured in the laboratory. This necessitates removing samples from the kiln. A minimum number of six is usually taken to assure representative sampling. The samples must accurately preserve a record of the orientation of the structure for the vector components to be measured as they occur *in situ*. This requires that the orientation of a plane in space must be established and that this plane shall have a fixed relationship to the sample. A square mould is fitted around the sample and filled with plaster, thus capping the sample. The top surface of the mould is made horizontal with a precision spirit level and this surface of the plaster becomes the reference plane. A bearing line is scribed on to this plaster surface.

FIG. 3. Measuring a sample in the laboratory

The remanent magnetism of the sample is measured with an astatic magnetometer; Helmholtz coils cancel the local magnetic field, thus eliminating induced magnetism. A sample, enclosed in plaster, is being positioned in its holder.

The bearing line is obtained by back sighting from the sample to a theodolite station with the aid of a telescopic alidade or similar device. The theodolite station is the reference point for all samples from one collecting site. From this station a Sun bearing, or triangulation by resection, can establish geographic north without recourse to any magnetic measurement. (Fig. 2.)

The plaster cap over and around the sample not only provides a horizontal surface but, by being square, four orthogonal vertical surfaces, which can be used as reference planes in the measurement of the remanent magnetic components. Furthermore, the plaster protects the samples which are sometimes quite fragile. An alternative scheme for orientation, recently described by Bidgood and Harland (1959), uses a miniature theodolite head which is fitted directly on the sample by securing it in three small drill holes.

Measurement of the remanent vector is made with an astatic magnetometer. The deflection of the magnet system when the sample is placed near it is noted from a distance by the familiar lamp, mirror and scale arrangement. Both the sample and the magnet system are placed in the central region of a Helmholtz coil system which cancels the Earth's magnetic field. This removes any large, induced magnetism from the sample, which might affect the measurement. The coils also permit measurement of the induced magnetism in weak magnetic fields when this information is desired. The coils are powered from accumulators and the current is monitored by a potentiometer circuit. (Fig. 3.)

It is assumed that a centred dipole approximates the magnetism of the samples. Six components of the remanent vector are measured. These correspond to the projection of a central vector on the sides of a cube enclosing it. The mean resultant vector is calculated from these components. The vector is rotated into its geographical attitude according to the orientation marks on the sample. Finally, the mean vector for the kiln is calculated from the individual resultant vectors and an estimation of the error is made following the method proposed by Fisher (1953) for dispersion on a sphere. A stereographic projection is helpful in plotting and manipulating these vectors.

When the triangles of error of the six vector components are large, some additional study is required. Inhomogeneity of the sample due to differences in the extent to which the clay has been fired, may make the centred dipole assumption quite false. Quartering the sample by means of a diamond saw and independent measurement of each sub-sample usually removes this limitation. In cases where instability of the magnetism is suspected, thermal or alternating electric current demagnetization experiments are carried out.

Though in these ways some additional results may be obtained, the results to be discussed here have no such material included in them. The experience has been that, with over half the material collected, the first measurements are useful. About 20 per cent has been useless.

The measurement of a single sample, including all corrections for orientation, etc., can be done to an accuracy of 1° in declination and slightly better than this figure for the inclination. The probable error for the mean of the samples from one kiln varies from less than one degree to four degrees for these results.

(c) Results

Through the spring of 1959, 87 structures had been sampled and measured in Britain. Twenty-four of these were done before the techniques had reached the final

development and precision discussed above. Sixteen were from structures other than kilns and had interests different from the secular changes we are discussing. Of the remaining 47 kilns, the results for 24 are shown in Fig. 4. The mean vector directions are shown and their approximate ages indicated where these are known. Eighteen of the samples fall in the Roman period and the remainder in the Mediaeval period. The paucity of material available from the period A.D. 500 to 1000 is making it difficult to obtain as complete a succession in that period as is desired.

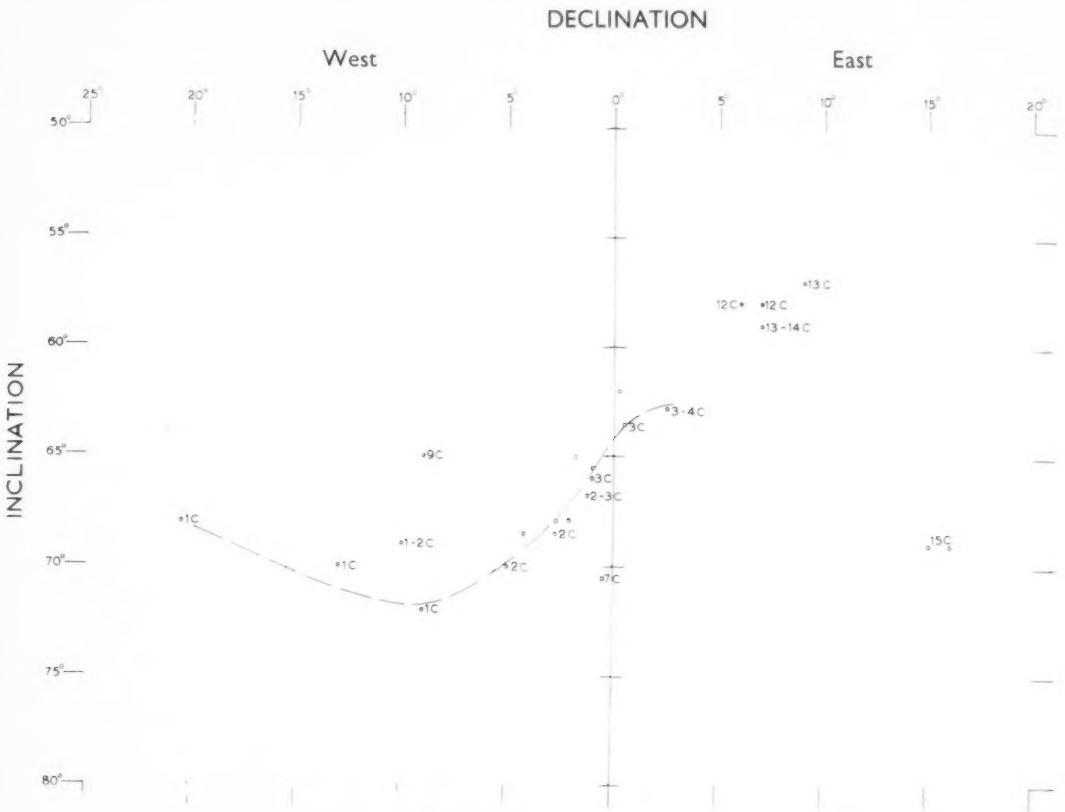


FIG. 4. Archaeomagnetic results for the Christian era in England. (□ = a result from Greece: see text)

These points, taken throughout England, but concentrated along the East coast, have been reduced to London for comparison. The method of doing this has been to apply the dipole field gradients applicable to a uniformly magnetized Earth whose field coincided with the measured vector. The latitude correction of the inclination is the largest factor in such a reduction.

The results are consistent with a pattern of change suggested by the curve sketched by dashes in Fig. 4. This curve applies to the 1st to 4th centuries A.D. After this time, and before observational records can be called on, the change is still not well defined, but what points have been determined are highly suggestive of the pattern. The

field appears to have been westward in its declination in the 9th century and then varied to an easterly declination and shallower inclination by the 12th century. By the end of the 13th century the change of inclination altered its sign and the declination and inclination increased until the 15th century. Very near the time of the first measurements of the magnetic declination in Europe, the easterly declination may have been near its maximum value in England.*

Before the 9th century and after the 4th, the record is still missing in England. Though efforts are being made to close this gap we must turn to a result from Greece for help at the present time. This is shown as a separate symbol in Fig. 4. The measurements are from a well-preserved kiln of the 7th century found near the Agora in Athens. This point suggests that the change traced out a curve with a clockwise sense during the period with which we are concerned. This result was extrapolated to London in the same manner as described for the English results. Such an extrapolation is not trustworthy and must be checked by the examination of materials of similar age in countries lying between England and Greece.

3. Other Indirect Methods

The study just outlined is an experiment to determine the geomagnetic secular variation of the past by the examination of fired structures collected *in situ*. Another aspect of archaeomagnetism which might be worth investigation is the study of the remanent magnetic inclination in pottery. Careful selection of the pottery type would be necessary for the assumptions concerning its orientation in the kiln to be simple and valid. Figured and glazed decorative or ceremonial pottery would be the general type of choice and one would assume that the axes of the pots had been vertical. (It can be demonstrated that this is not true of coarse, unglazed pottery.) Such pottery is distinctive, permitting a careful division as to age and region, and is already well represented for many periods in museums, permitting adequate sampling to avoid random errors. A step in this direction had been indicated by Aitken (1958) who reported measurements on Chinese Yüeh ware of the 10th and 12th centuries.

Leaving these anthropocentric studies, one could enquire into natural processes which might record the geomagnetic field repeatedly and precisely enough to permit the study of the secular variation. Two rock-forming processes suggest themselves as meeting these requirements: the slow accumulation of fine sediment by gravitational settling in still water and the cooling of small igneous rock bodies from temperatures above their Curie points.

One example of sedimentation magnetization which has been well investigated is that found in varved clays. These clays, which exhibit a layering attributed to the year's seasonal changes, were laid down in small basins of deposition near the ice front during the periods of glaciation. Their remanent magnetism was first investigated by Ising (1943) and McNish and Johnson (1938). But encouraging though these results were, they did not result in a continuous and accredited record of the secular variation, and the recent work of Griffiths, King and Wright

* Speculation concerning the influence this had on the rejection as a result of experimental observations of the notion that the compass pointed to the Pole Star has some justification. It appears that the compass then pointed further from geographic North than at any other time since it had been introduced in Europe.

(1957 and 1958) suggests these results must be viewed with scepticism. The disrupting influence of water currents in the basins of sedimentation seems to produce a false alignment of the small magnetic particles.

Another environment where the sedimentary particles might experience only small mechanical forces and become aligned with the magnetic force is the deep ocean. An investigation of some marine cores was reported by Johnson, Murphy and Torrenson (1948). They found variations in the direction of remanent magnetism comparable to the change in the magnetic field which man has observed. Work of a similar nature has been begun at Cambridge University by M. J. Keen (Report of a discussion, 1959). One difficulty arising in the abyssal sediments is that the sedimentation rate is very slow. The time span covered by a specimen of the size necessary for accurate measurement is likely to be greater than 1000 years. Again the disruption due to marine currents must be considered (Heezen, 1959) and also the effect of burrowing animals (Laughton, 1957). Results must be gathered for two or three closely spaced cores and repeated for a proper sampling of the principal types of abyssal sediment before this study can be properly assessed.

The thermo-remanent magnetism of some lava flows appears uniform enough to constitute a single precise observation with respect to time of the geomagnetic field. Coevallier (1925) attempted to use this method of documenting the secular variation by sampling historically recorded lava flows at Mount Etna ranging from the 12th to 17th centuries. Though most lavas cannot be dated with great precision, it can be hoped that enough results will be gathered from dateable flows to add considerably to the secular variation record.

Some intrusive igneous rock bodies are large enough for their cooling history to extend over times of the order of 10^2 to 10^3 years. It would be expected that different parts of such bodies would pass through the Curie temperature at quite different times and thus acquire different directions of magnetization. Jaeger (1957) discussed this cooling problem and then attempted to apply it in a study of the secular variation by measuring several samples spaced out across a sill in Tasmania which is 430 metres thick (Jaeger and Green, 1956). Unfortunately the accuracy of the measurements was not sufficient to establish that a non-uniform magnetization was characteristic of the body. Though this method cannot yield absolute values for the variation period, it has considerable potential worth, if established, for it can be applied far back in geological time. It might tell if the rate of change had varied by large amounts during the Earth's history and what were the maximum deviations from the mean of the magnetic directions.

4. Conclusions

Certain conclusions may be drawn from the results reported. In the Roman period, the rate of the secular variation in direction of the magnetic force was very similar to that of the period of direct observation. The view that a true periodicity occurs in this variation is not supported and extrapolations back in time based on this assumption are almost certainly false (see Watanabe 1958). Throughout the period this work has covered (since about A.D. 50) the mean direction of the magnetic force appears to have been some degrees West of geographic North.

This last observation should be taken as evidence against the notion that the secular variation is simply related to a revolution of the Earth's magnetic poles

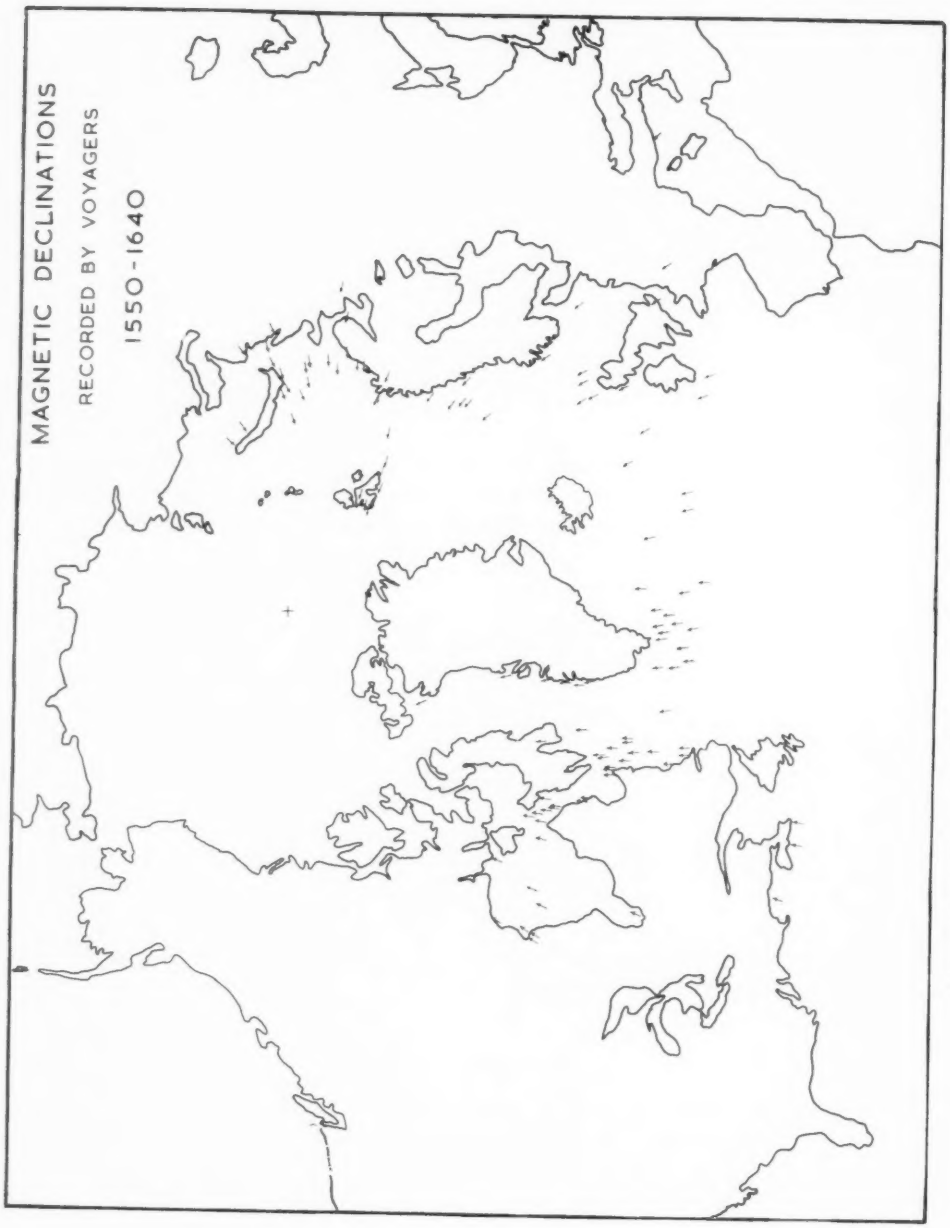


FIG. 5. Declinations recorded by voyagers during the period A.D. 1550–1640.

VOL
4
196

around the geographic poles. This is no longer a widely held belief but is occasionally encountered. Bullard *et al.* (1950) have shown that there has been no evidence of any movement of the geomagnetic pole since 1880. Information from earlier periods is inadequate to well define the geomagnetic pole by a harmonic analysis of the magnetic field. Results of the sort described in Section 2 may permit a determination of the ancient dip poles (commonly called the magnetic poles) when a sufficient number of the Earth's regions have been studied to allow the declination to be used in triangulating the pole positions for certain periods of antiquity. For the moment, with results only from Europe, this is not possible. However, records do exist which permit this for the close of the 16th century. These are the records of magnetic declinations made by explorers. Such people as those searching for northern sea routes to Asia made many determinations. Many of these have been tabulated by van Bemmelen (1899). This is the source from which the author has prepared Fig. 5. It can be seen on this map that over much of the northern hemisphere the compass pointed to the same area as it does today. We can conclude that there has been no large movement of the magnetic pole in the northern hemisphere during the last four centuries.

It was only in northern Europe that declinations quite different from this norm were recorded at that time and these are shown in Fig. 5. This is another expression of the easterly declinations shown for London during that period in Fig. 1. Obviously a magnetic disturbance seems to have affected that region. Van Bemmelen (op. cit.) produced charts showing a westward drift of this disturbance in the 16th and 17th centuries. The centre crossed northern Europe from Scandinavia and finally died out in the Atlantic by early in the 18th century. These charts of van Bemmelen cannot be considered to be very exact, as the information from which they were compiled was very scanty and varied much in its accuracy. However, by now they have historic interest in the study of geomagnetism and do graphically present a disturbance which was well documented in Europe. As the charts are not easily accessible to many students, two, those for 1600 and 1700, have been reproduced as Figs. 6 and 7.

We have, then, some record of a large magnetic disturbance which affected magnetic observations in Europe in the late medieval period. This disturbance died out rapidly during the 17th century. The results given in Section 2 show that this disturbance developed between the 9th and 12th centuries and appears to have changed rather slowly between the 12th and 16th centuries. Much remains possible to do in better documenting this feature.

Such knowledge of the growth and decay rates of the centres of secular magnetic change may be of help to theories concerning the Earth's main magnetic field, the physical nature of the motion producing the disturbance, and the magnitude of the magnetic fields acting within the Earth's core. Thus extending the work in space as well as time, to other continents, is important. We would like to know if the low rate of geomagnetic secular variation which has been observed in the Pacific area is a permanent feature of the magnetic field or a temporary feature due to the random distribution of the few centres of secular change now present on the Earth. The cultural histories of China and India make those regions particularly suitable for investigations of the sort described in Section 2.

This description and discussion of the archaeomagnetic studies of secular variation has stressed its geophysical importance. There is a direct application of this work

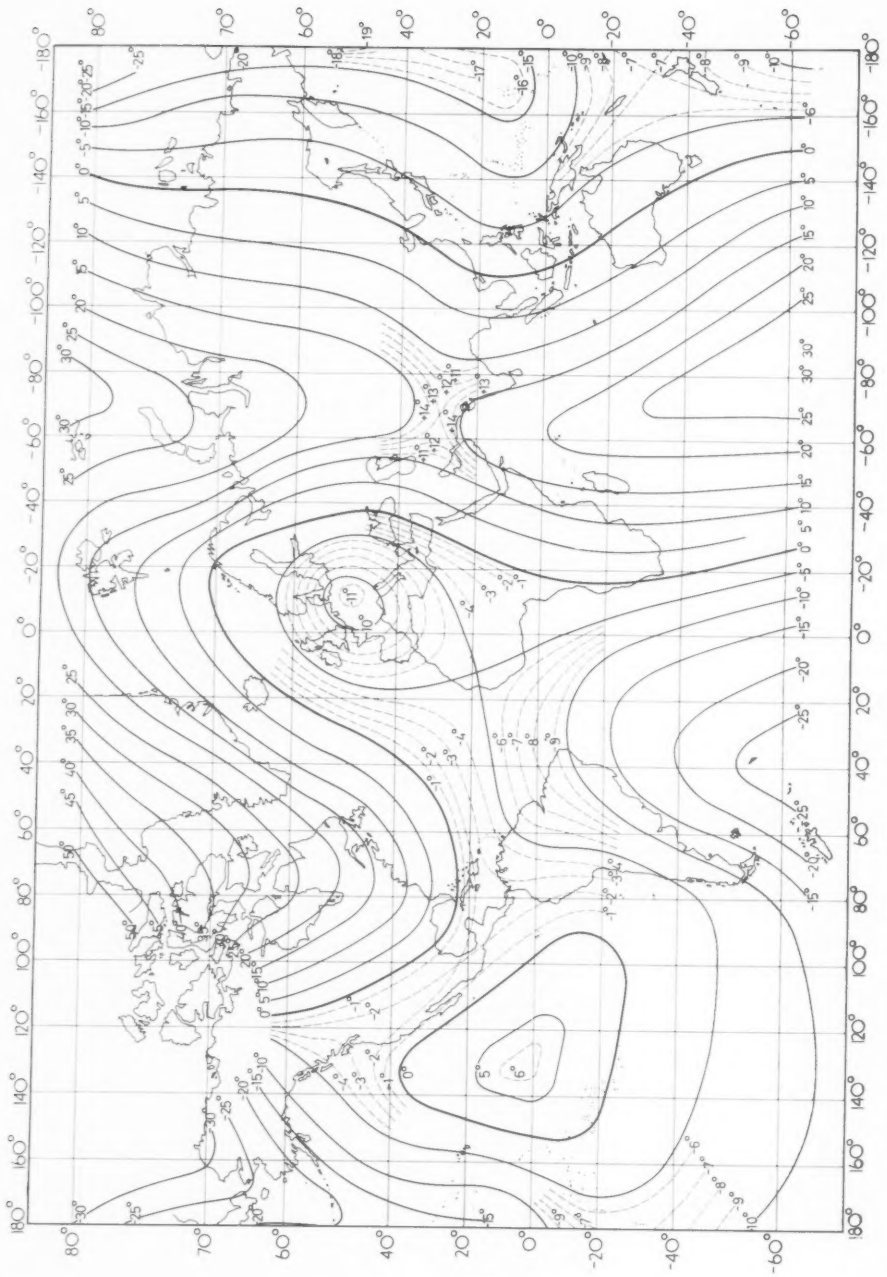


Fig. 6. Isogonic lines for A.D. 1600 (after VAN BEMMELEN).

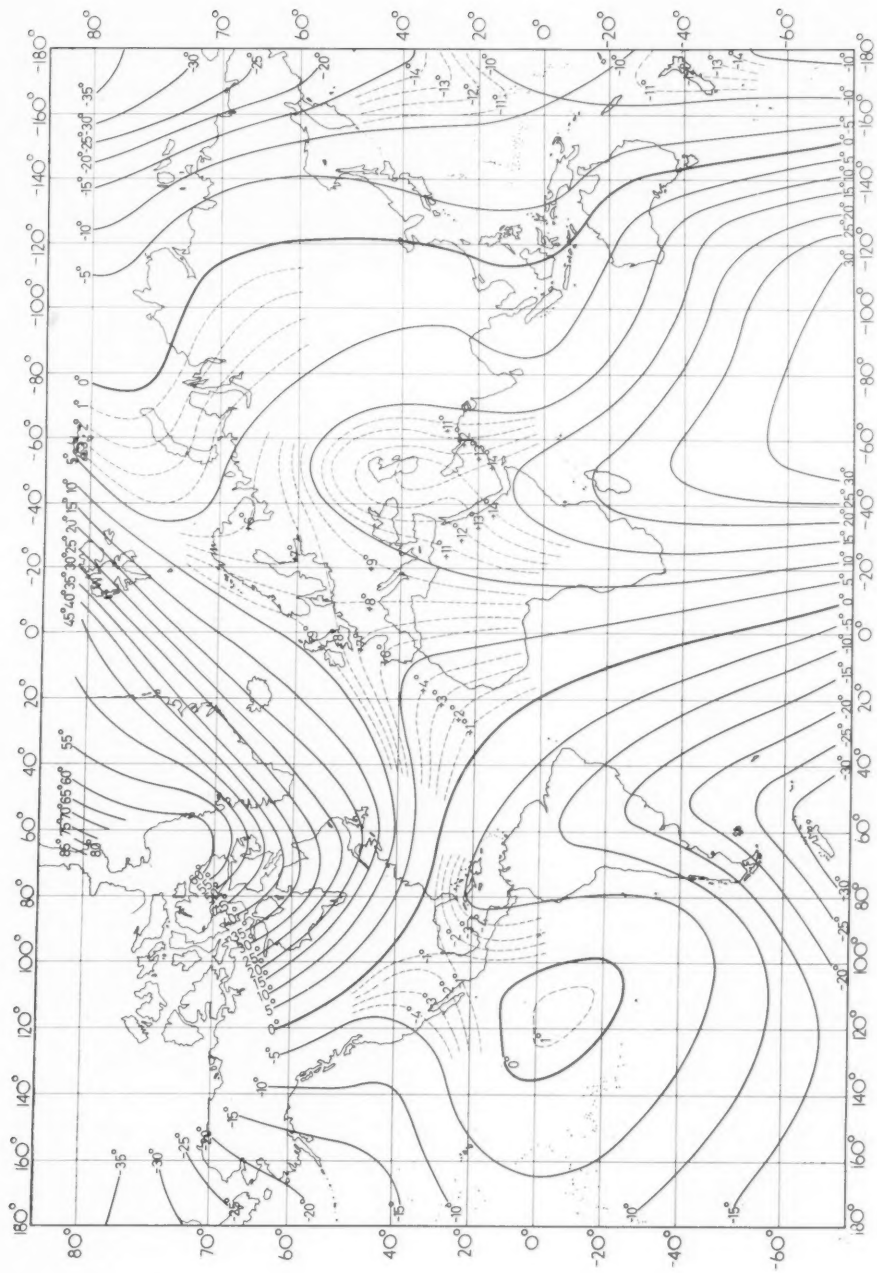


FIG. 7. Isogonic lines for A.D. 1700 (after VAN BEMMELEN).

which should also be recorded. This is in the use of the magnetization of the kilns to date them. Working from the curve established for sites where the dates are well known on other grounds, the chronological order of a sequence of sites becomes clear.

The regularity of the curve of secular change ensures that relative dating between sites should be easy. Absolute dating is dependent on the dating of the sites by which this empirical curve is anchored. Historic record is the most useful aid back through the classical period. In Section 2 an accuracy of $\pm 1^\circ$ was said to be attainable. At the present rate of magnetic change (about 1° in six years at Cambridge) an accuracy to about a decade can be envisaged under the most favourable conditions. This is often sufficient to permit discrimination between the ages of sites which are approximately dated on the basis of their associated objects found during excavation.

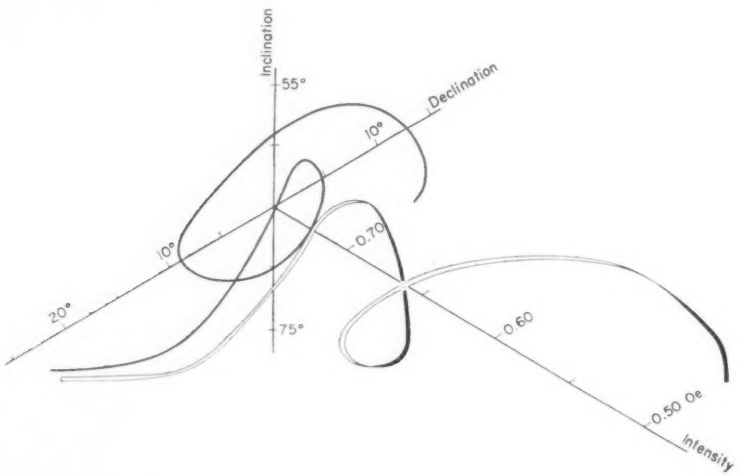


FIG. 8. Isometric presentation of a curve relating the secular variation in declination, inclination and total intensity of the magnetic force.

As the curve of change has no simple analytic expression, the possibility of ambiguity due to the curve crossing itself, is quite real. At least one well established point for each century during the complete period covered by the study, should be the aim of workers developing this method. Though such a method is empirical and can be applied only after a detailed study has already been made, one point should be noted: the rate of change appears to have been reasonably constant and hence the accuracy of this method would not vary greatly between materials of different ages.

No mention has been made of secular change in the intensity of the magnetic field. THELLIER (1959) has particularly concerned himself with this problem. His results suggested that the intensity has declined steadily throughout the Christian period from a value in the first century A.D. about 70 per cent greater than the present one. This is in agreement with the findings of magnetic observatories during the past century which show a 5 per cent decline in that time. A study of this sort requires careful checks that the laboratory experiments which are made to establish the strength of the magnetic field acting on a sample when it was magnetized do not alter the magnetic nature of the sample. Also, care must be taken to ensure that one is properly comparing materials from scattered sites as the variation in intensity

which occurs with geomagnetic latitude is quite large. THELLIER appears to have shown that these problems are not insurmountable and we can only now hope for many such determinations to amplify and extend his extremely interesting results. In Fig. 8, the variation has been assumed to be linear in time and the secular change of direction and intensity is shown as an isometric drawing. This curve suggests how knowledge about the intensity of magnetization of samples might remove ambiguities in the curve of directional change and be of direct use in the application of archaeomagnetism to problems of chronology.

The results which have been discussed are only those of the early progress of a study which may require several years before the secular variation of the geomagnetic field in Europe is documented for the last 3000 years.

REFERENCES

AITKEN, M. J. . . . . 1958 Magnetic Dating I, *Archaeometry*, **1**, 1, 12–20.

BIDGOOD, D. E. T. and HARLAND, W. B. 1959 Rock Compass: a new aid for collecting orientated samples. *Bull. Geol. Soc. of America*, **70**, 641–4.

BOROUGH, WILLIAM . . . . . 1580 Declination measurements, referred to by ROBERT NORMAN, London 1581, whose paper is reprinted in "*Rara Magnetica*", selected by G. HELLMANN as No. 10 of his "Neudrucke von Schriften und Karten über Meteorologie und Erdmagnetismus" publ. by A. Asher & Co., Berlin, 1897.

BULLARD, E. C., FREEDMAN, C., GELLMAN, H. and NIXON, J. 1950 *Phil. Trans.* A**243**, No. 859, 67–92.

CHAPMAN, S. and BARTELS, J. . . 1940 (Reprinted 1951). *Geomagnetism* (2 vols.) Oxford, 156–7.

CHEVALLIER, R. . . . . 1925 L'amantation des laves de l'Etna et l'orientation du champ terrestre en Sicile du XII^{eme} au XVII^{eme} siècle. *Ann. de Phys.*, 10eme série, 4.

COOK, R. M. and BELSHÉ, J. C. . . 1958 Archaeomagnetism; a preliminary report on Britain. *Antiquity*, **32**, 167–78.

ELSASSER, W. M. . . . . 1955 Hydromagnetism. A Review. Part I. *Amer. J. Phys.*, **23**, 9, 590–609.

ELSASSER, W. M. . . . . 1956 Hydromagnetism. A Review, Part II. *Amer. J. Phys.*, **24**, 2, 85–110.

FISHER, R. A. . . . . 1953 Dispersion on a sphere. *Proc. Roy. Soc.*, A**217**, 295.

GRIFFITHS, D. H., KING, R. F. and WRIGHT, A. E. 1957 Some field and laboratory studies of the depositional remanence of recent sediments. *Adv. Phys.*, **6**, 22, 306–17.

GRIFFITHS, D. H., KING, R. F. and WRIGHT, A. E. 1958 An assessment of the difficulties involved in using quaternary varved sediments for palaeomagnetic studies of the secular variation. *Ann. Géophys.*, **14**, 4, 515–18.

HEEZEN, B. C. . . . . 1959 Dynamic processes of abyssal sedimentation in erosion, transportation, and redeposition on the deep-sea floor, *Geophys. J.*, **2**, 2, 142–63.

ISING, G. . . . . 1943 On the magnetic properties of varved clays, *Ark. Mat. Astr. Fys.*, I, **29**, A No. 5.

JAEGER, J. C. . . . . 1957 Temperature in the neighbourhood of a cooling intrusive sheet, *Amer. J. Sci.*, **255**, 4, 306–18.

JAEGER, J. C. and GREEN, R. . . 1956 The use of the cooling-history of thick intrusive sheets for the study of the secular variation of the Earth's magnetic field, *Geofisica*, **35**, 3 49–53.

JOHNSON, E. A., MURPHY, T. and TORRENSON, O. W. 1948 Pre-history of the Earth's magnetic field, *Terr. Mag.*, **53**, 349–72.

KEEN, M. J. . . . . 1959 *Geophys. J.*, Vol. 2, p. 170.

LAUGHTON, A. S. . . . . 1957 Exploring the deep ocean floor, *J. Roy. Soc. Arts.*, **106**, 39–56.

MCNISH, A. G. and JOHNSON, E. A. . 1938 Magnetization of unmetamorphosed varves and marine sediments, *Terr. Mag.*, **43**, 402–7.

Report of Discussion	1959	The geological and geophysical results of the cruises of R.R.S. *Discovery II*, 1958, May–July, *Geophys. J.*, **2**, 2, 168–70.
Thellier, E.	1938	"Sur l'aimantation des terres cuites et ses applications géophysiques". Thesis Paris; and in *Ann. Institut de Physique du Globe, Paris*, **16**, 157–302.
Thellier, E. and Thellier, O. . .	1959	Sur l'intensite du champ magnetique terrestre dans le passe historique et geologique, *Ann. de Géophys.* **15**, 3, 285–376.
van Bemmelen, W.	1899	Die Abweichung der Magnetnadel . . . bis zur Mitte des 18. Jahrhunderts, *Batavia R. Magn. Obs.*: Supplement to Vol. 21 of *Observations* 1899.
Vestine, E. H., Laporte, L., Lange, I. and Scott, W. E.	1947	The geomagnetic field, its description and analysis, *Carnegie Instn. of Washington*, Publ. 580.
Vestine, E. H., Laporte, L., Lange, I., Cooper, C. and Hendrix, W. C.	1947	Description of the Earth's main magnetic field and its secular change, 1905–1945, *Carnegie Instn. of Washington*, Publ. 578.
Watanabe, T.	1958	*J. Geomagn. and Geoelectr.* **10**.

VOL
4
196

The Development of a Solar Centre of Activity

C. DE JAGER

Sterrewacht Sonnenborgh, Utrecht, Netherlands

1. INTRODUCTION

THE life-history of a typical solar Centre of Activity, ("CA"), was first described by KIEPENHEUER (1953). Since then our knowledge of activity centres and of the phenomena taking place in them has considerably increased. KIEPENHEUER's original description, however, remains essentially valid, and it will serve as a basis for the present discussion. Nevertheless, it needs to be supplemented in a number of very important details. Furthermore, during recent years the theoretical understanding of many features has also much improved.

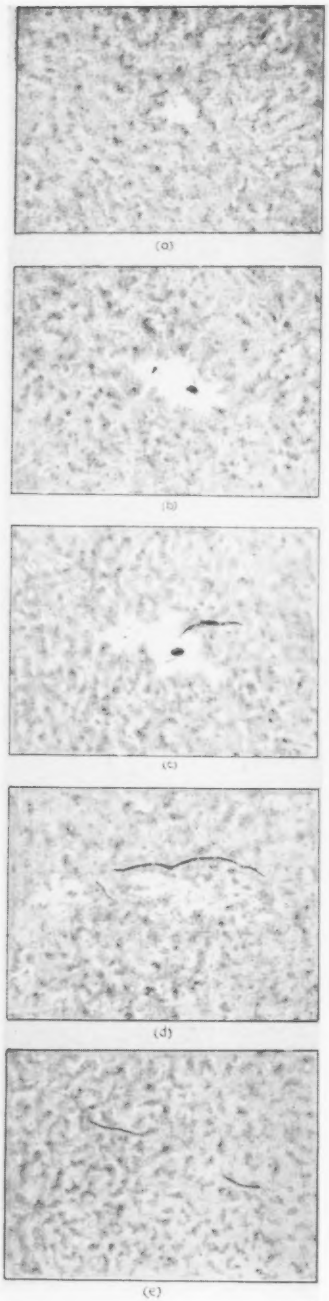
A solar Centre of Activity is a limited and an originally coherent region of the Sun; it has an area usually smaller than about one tenth of that of the visible solar hemisphere. It is in these centres that the transient phenomena occur, the spots, faculae, flares, surges, prominences, coronal condensations. It presents a bipolar magnetic field. The enhanced solar radio-noise, the enhanced X-ray radiation, and perhaps also the solar cosmic rays, are emitted by these regions.

A number of these CA's may occur on the disc simultaneously, although they originate at different moments and have different lifetimes and life histories. Many of them are short-lived, they exist sometimes only for a few weeks, and never reach appreciable dimensions. Only a few of them live for a long time, say, 100–200 days, and sometimes even longer. These phenomena are said to pass through a "complete development" and, although rarer than the short-lived ones, they are generally described as the "typical CA's"; this does not indicate that most CA's resemble them, but only that they are considered as a kind of ideal, or, at a certain stage, as fully developed CA's.

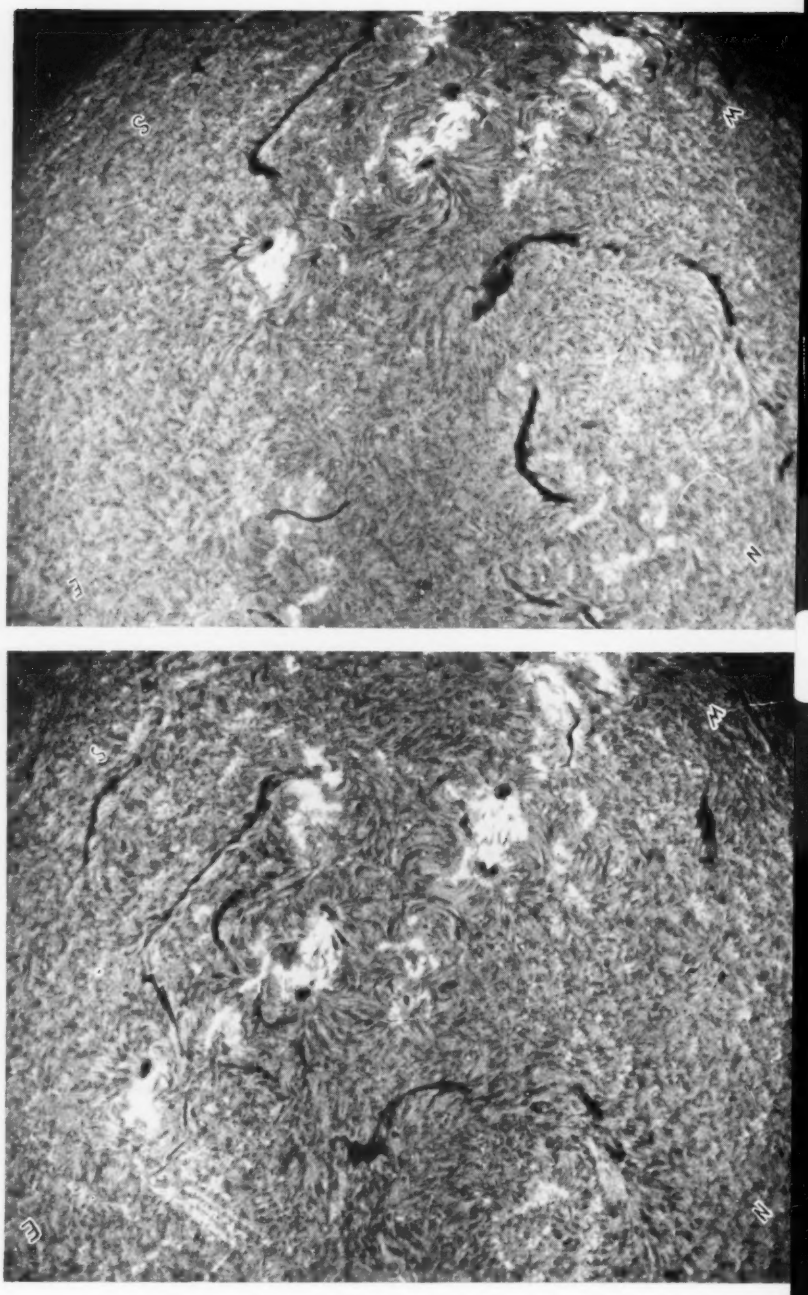
In this introductory Section we shall give a short review of the phenomena occurring in such a "typical CA". In the subsequent Sections some of the details will be worked out, and an attempt made to synthesize into a single picture the different observations (which relate to different parts of the CA and are obtained with different instruments). A complete synthesis is not yet possible at this stage; hence this paper must be considered as a progress report, marking the present stage of knowledge in this fascinating field of research.

We shall not enter into a detailed discussion of any of the instruments or of the observational methods, and shall assume that their mode of functioning is known.

The Development of a Typical CA, as seen on an Hα or Ca^+ spectroheliogram, is shown by the sets of Figs. 1, 4 and 11, the first of which is reproduced from those originally given by KIEPENHEUER. Roughly, the life-history of a typical CA may be divided into four parts:



(a)



(b)

FIG. 1 (*a*) and (*b*)

(*a*) The development of chromospheric structures, viewed in Hα, in a Centre of Activity; (After KIEPENHEUER, [1953] slightly modified. See text).

(*b*) Changes in Activity Centres. The Sun photographed at the Anacapri station of the Freiburg Observatory (through an Hα filter). 1958 December 6 (13 hr 06 m'n) and 8 (14 hr 04 min). South is top left; West is top right. (Courtesy K. O. KIEPENHEUER.)

1. Pre-spot phase: development of a bipolar magnetic-field region, (BM), followed by facular fields, still without spots. The limits of the faculae coincide more or less with those of the magnetic-field region. We note the formation of a coronal activity region. Duration of this phase: one to several days. [Fig. 1 (*a*) (a).]

2. Spot phase. This is the most important phase. The faculae and the magnetic-field region increase in size. Spots develop, spot prominences and flares and their associated phenomena, such as surges and sporadic condensations, etc., occur. Many of the solar radio-bursts are emitted by the CA at this phase, and enhanced X-ray radiation and (sometimes) cosmic rays. Duration of this phase: one to several weeks [Figures 1 (*a*) (b) and 1 (*a*) (c).]

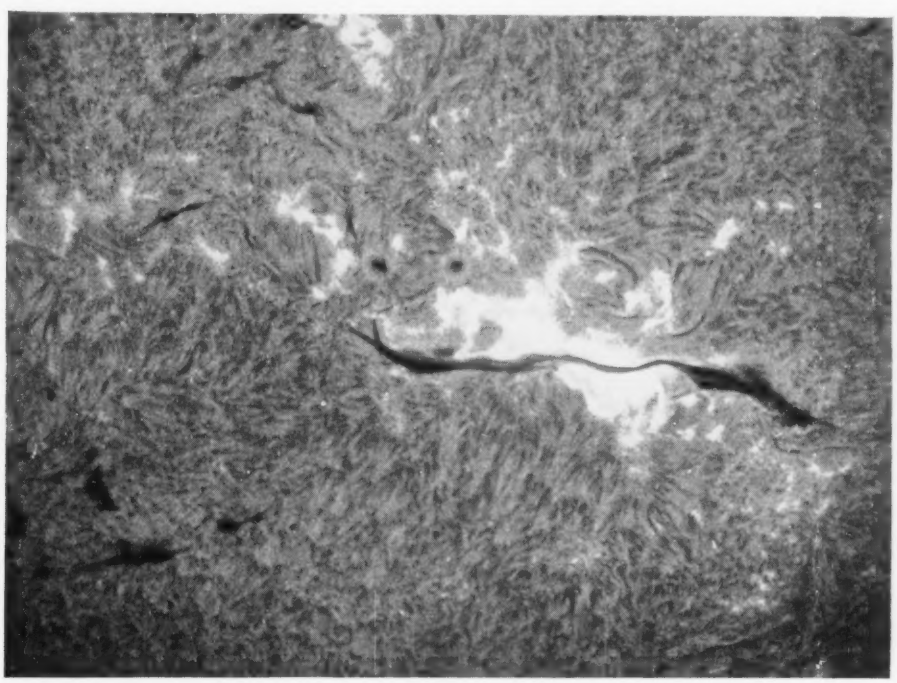


FIG. 1 (*c*)

(*c*) Filaments and Activity Centres. Photograph, 1957 November 13, obtained at the Anacapri station of the Freiburg Observatory. (Courtesy K. O. KIEPENHEUER.)

3. Bipolar Magnetic (BM) post-spot phase. The spots, flares and associated phenomena have disappeared; there remains a magnetic bipolar facular region that increases in area; it gradually dissolves and then decreases in intensity. The long-lived quiescent filaments (and perhaps the coronal rays) are associated with this phase. Duration: one to several months. [Fig. 1 (*a*) (d).]

4. These phases might be followed by the Unipolar Magnetic (UM) phase. The magnetic field has much decreased in intensity and has become unipolar; UM-fields are very rare and have only been observed around 1954 during a period of low solar activity. No optical signs of activity are left. Duration: months.

These four phases will be discussed below, with the second phase receiving most of our attention.

2. Description of the First Phase of a CA

Magnetograph observations of the Sun, which are now in progress at the Mt. Wilson, Crimean and Cambridge observatories, have revealed that the development of a CA is intimately related to the magnetic behaviour of the solar surface. In the first instance, before any optical phenomenon becomes visible, a magnetic field originates. After a short lapse of time (some hours to some days, in rare cases a fortnight) this is followed by optical phenomena in the following sequence:

1. One of the dark mottles, visible on an Hα spectroheliogram, elongates and gets brighter, becoming a facular point with an Hα central-intensity about half that of the adjacent continuous spectrum. Sometimes even two bright mottles originate, which are connected by a dark "flocculus". Spots develop inside these mottles, often after some hours and sometimes, after a longer interval, of one to several days. The magnetic-field region increases in area, and takes on a bipolar character, in line with the bipolar character of the spot-group.

2. In the corona, above this growing CA, one observes, simultaneously with the photospheric and chromospheric phenomena, a decrease of the coronal line radiation at 6374 Å (Fe X; $\chi = 233$ eV) and an increase of the radiation at 5303 Å (Fe XIV; $\chi = 357$ eV). This indicates, perhaps, a slight increase of the coronal temperature, since the ionization of Fe XIV and the excitation of the 5303 Å line demands a temperature of about 1×10^6 °K, somewhat greater than the ionization temperature of 7×10^5 °K assumed for the quiet corona. At the same time the density in such a coronal activity region increases to about twice the normal coronal level, as is indicated by the increase of the radio-radiation at decimetre wavelength, observed in regions above faculae, and by intensity measurements in the K corona (see Section 5).

3. The Magnetic Field of a CA

Our discussion of the observations, which were summarized in the preceding section, consists of two parts. First, it is necessary to investigate the type of mechanism which is responsible for the facular field, the sunspots and the coronal activity centres; this is done in the present Section 3. Next, (Section 4), once this mechanism has been found, the nature of its driving forces should be investigated.

There are good reasons for assuming that the fundamental phenomenon within the CA in its first phase is the *magnetic field*. By some unknown cause a magnetic field arrives at photospheric and higher levels and produces the observed phenomena. The subsequent increase of area of the CA must be a result of the increased area of the magnetic-field region. The observations show that at first a weak field appears, which leads to a local increase of the brightness of the chromosphere, best visible in Ca^+ spectroheliograms but also in Hα. Then, a sunspot forms in this field. In contrast to the *bright* chromospheric and photospheric faculae, this is a *dark* region of the photosphere and of the lower chromosphere. However, it often happens that far above the spots, in chromospheric images obtained in Hα, the spots are no longer visible, being covered by bright faculae with a brightness similar to that of the normal facular field: so, far above a spot, the matter emitting Hα becomes *bright*.

Essentially, a spot is a rather sharply limited concentration of magnetic lines of force in a region of a weaker field (the BM region). The spot's magnetic fields may range between some hundreds and some thousands of gauss, whereas the faculae

have an average field of some tens of gauss, which locally may attain values of some hundreds of gauss (cf. Fig. 8 (*b*)). The fact that the photosphere is dark in the places of spots is due to the magnetic *inhibition of convection* in that region (BIERMANN, 1941). Convection, which is the main mechanism of energy transport in an extended solar region below the undisturbed photosphere, occurs at a very reduced rate in the spot region, thus locally reducing the energy flow, and causing the darkness of the spots.

Inhibition of Convection. (See also COWLING, 1933, 1957; concerning the structure of sunspots see also SWEET, 1956; MAKITA, 1958). If convection is defined as systematic up- and down-ward motions initiated by thermal instability, the common statement that convection is inhibited by the magnetic fields of spots does not seem to be wholly true. Laminar motion in the direction of the field is certainly possible, but motions perpendicular to this direction would be damped.

In matter moving with the velocity $\mathbf{v}$ in a field H, the electric current j is defined by

$$\boldsymbol{j} = \sigma(\boldsymbol{E} + \mathrm{c}^{-1}\mathbf{v} \wedge \boldsymbol{H})$$

where σ is the electrical conductivity; we put the magnetic permeability $\mu = 1$. Since, further, curl $\boldsymbol{E} = -\mathrm{c}^{-1}\partial\boldsymbol{H}/\partial t$, curl $\boldsymbol{H} = 4\pi\boldsymbol{j}\mathrm{c}^{-1}$ and div $\boldsymbol{H} = 0$ we have

$$\partial\boldsymbol{H}/\partial t = \frac{1}{\sigma}\,\mathrm{curl}\,(\mathbf{v}\wedge\boldsymbol{H}) - \frac{\mathrm{c}^2}{4\pi\sigma}\nabla^2\boldsymbol{H}. \qquad \ldots\ (1)$$

Defining λ and τ as a length and a time, respectively, within which the field changes by a factor of the order of unity, equation (1) becomes

$$\frac{H}{\tau} \approx \frac{\mathbf{v}H}{\lambda\sigma} - \frac{\mathrm{c}^2 H}{4\pi\sigma\lambda^2}. \qquad \ldots\ (2)$$

Writing $v = 2{\cdot}10^5$ cm/sec, $\lambda = 5{\cdot}10^7$ cm, $\sigma = 10^{13}$ e.g.s., (values valid for the upper part of the convection region). Equation (2) shows that $v/\lambda \gg \mathrm{c}^2\,(4\pi\sigma\lambda^2)^{-1}$; hence Equation (1) reduces to

$$\partial\boldsymbol{H}/\partial t = \mathrm{curl}\,(\mathbf{v}\wedge\boldsymbol{H}). \qquad \ldots\ (3)$$

This equation expresses that the motions of matter and of the magnetic lines of force are intimately connected; they move together. The induced electrical force appears to be negligible as compared to the field changes due to the material movement. Any material motion $\mathbf{v}$ perpendicular to the field will induce a current $\boldsymbol{j} = -\mathrm{c}^{-1}\sigma\mathbf{v}\wedge\boldsymbol{H}$, and this leads to an opposing force

$$\boldsymbol{F} = \rho\,\mathrm{d}\mathbf{v}/\mathrm{d}t = \boldsymbol{j}\wedge\boldsymbol{H} = -\mathrm{c}^{-1}\sigma\mathbf{v}H^2 \qquad \ldots\ (4)$$

So the motion $\mathbf{v}$ will be damped exponentially in a characteristic time $\tau = \mathrm{c}\rho/\sigma H^2$; at the solar surface we have $\sigma = 10^{13}$ e.g.s.; $\rho = 4{\cdot}10^{-7}$ g.cm.$^{-3}$ $\tau = 50/H^2$ sec; hence τ is very short in a sunspot. The lines of force are said to be "frozen into the matter". Since, however, the magnetic energy-density $H^2/8\pi$ in the upper part of the convection zone exceeds the kinetic energy of the convective motions, the field motion determines the material motion. Hence, in that region, an ascending motion of a convective element cannot be accompanied by an (adiabatic) expansion, and so the ascending element rapidly becomes heavier than its surroundings and sinks: convection is inhibited.

The lower limit H_l of the magnetic field at which inhibition occurs, can be derived from the equality between the energy of the convective motions $\frac{1}{2}\rho v^2$ and the magnetic energy-density $H^2/8\pi$. It appears that H_l depends on the depth in the convection zone as described below. The densities and velocities in the following table are taken from BÖHM-VITENSE's model of the solar convection zone (Vitense 1953, Böhm 1958); the molecular weight is $\mu \approx 1$:

Table 1

Depth (km)	Convective Speed v(km/sec)	log ρ	H_l (gauss)
200	1·0	−7·55	70
500	2·5	−6·4	500
1000	1·5	−6·0	600
2500	0·9	−5·15	1000
4500	0·6	−4·3	1500
7500	0·3	−3·35	2000

The upper limit of the convection zone is at about 100 km below the surface of the photosphere, but it is only from a depth of about 500 km and deeper inwards that convection provides the main part of the energy transport. This explains, in conjunction with the table, why an appreciable darkening of a spot can only occur in regions with magnetic fields greater than about 500 gauss. Regions with smaller magnetic fields should present themselves as "invisible spots", like those with field strengths up to several hundreds gauss, found by BEGGS and VON KLÜBER (1956) with their magnetograph.

It is also clear from the above table that most spots are rather shallow if only their coolness is considered. The field may penetrate deep into the convection region, but below a certain depth, where $H < H_l$, convection is not inhibited; hence the spot has the same temperature as the surroundings. The energy which is prevented from being transported upwards is for the greater part reflected downwards, and subsequently absorbed; it is thus leading to a local source of heat, which is situated in the convection zone below the region where we have $H = H_l$. The energy which radiates from this region suffices to explain the residual brightness of the umbrae of spots. The sideways-radiation may explain the "bright ring"; its excess brightness is 3%. It should not be confused with the faculae.

That the *faculae*, in contrast to the spots, are not dark but *bright*, shows that in the chromosphere, where convection is insignificant as a means of energy transport, the presence of relatively weak magnetic fields apparently produces a local heating of the relevant upper photospheric and chromospheric levels. Thus, the next problem is, why the weak fields of the order of some tens of gauss occurring in extended regions around the spots, should, unlike the spot-fields, be related to a *brightening* of the chromospheric and upper photospheric regions. This phenomenon is strongly related to the mechanism of the *heating* of the chromosphere; this heating is apparently greater in the facular fields than outside them.

The chromosphere is heated by a mechanical energy flux arising from the convective zone of the Sun. For the purpose of a mathematical description we note that the curl-free part of the velocity field of the convective zone is often represented by a field of progressing sound-waves, where the amplitude component of the sound waves may be put equal to the mean velocity component v_t of the photospheric

"turbulence". If there is no dissipation of turbulent energy, v_t increases outwards according to the relation: $\frac{1}{2}\rho v_t^2 \, . \, v_{\text{sound}} = \text{constant}$. When v_t becomes equal to the local velocity of sound, the sound waves are transformed into shockwaves. Observations of chromospheric turbulence show that we have $v_t \approx v_{\text{sound}}$ at a height of roughly 1000 km above the photosphere (DE JAGER, 1959a). This fact may explain why the fine mottling of the chromosphere only becomes visible in disc spectroheliograms taken in light emerging from heights above 1000 km: it seems that the splitting of the chromosphere into two kinds of constituents only takes place at a greater height (DE JAGER, 1959b).

The problem of how the mechanical energy of these shock-waves is dissipated and transformed into heat, giving a higher kinetic temperature of the chromospheric-gas, has been the subject of much controversy; the mechanism proposed by PIDDINGTON (1956, 1957) does not play any rôle as far as the heating of the normal chromosphere is concerned, contrary to what has often been assumed previously. For the heating of the faculae, however, PIDDINGTON's mechanism might be the most likely one: that the heating could be due to dissipation by *hydromagnetic* (*Alfvén*) *waves*, caused by frictional loss of energy between the oscillating ion-plasma and the neutral atoms. The latter tend to move independently, but they are forced to oscillate with the ions. This dissipation of energy is greatest for a field of some tens of gauss.

Alfvén-waves are important when the wave-velocity exceeds that of sound v_s (VAN DE HULST, 1951). Therefore, in order that Alfvén waves originate, the magnetic field H should be such that $v_A = H(4\pi\rho)^{-\frac{1}{2}} > v_s$. The table below gives the values H^*, when $v_A = v_s$. The interpretation of this table is that Alfvén-waves should become important in the fine mottles of the chromosphere as soon as $H > H^*$.

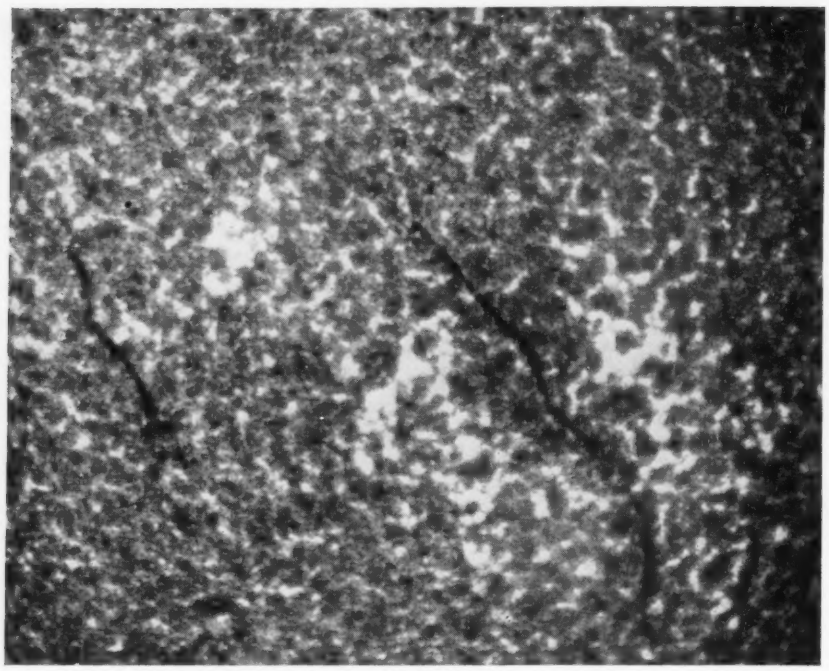
Table 2

Height above Photosphere km	log ρ	H^* gauss
0	− 7·7	500
1000	−10·1	30
2000	−12·3	25
4000	−13·0	1

Accepting the arguments given in the last part of this section we assume the field in the fine mottles to be of the order of 10–50 gauss. With the aid of the above table this explains why the chromospheric faculae are seen best in spectroheliograms taken at such a wavelength that the light emerges from chromospheric layers above about 1000 km (cf. DE JAGER, 1959a). Conversely, the fact that faculae are visible even in the low chromosphere and in high photospheric levels may either be an indication of the occurrence of high fields in the faculae proper, or that Piddington's mechanism does not yet wholly describe the dissipation process.

Thus the heating of the facular regions of the chromosphere seems to be a consequence of the magnetic field prevailing in these regions. A closer investigation of the *fine structure of the faculae* will provide more details concerning these regions of dissipation, and hence of the magnetic-field region.

The normal chromosphere when viewed on the disc is by no means homogeneous, but shows a structure which reveals three kinds of elements (DE JAGER, 1957, 1959a; see also BRUZEK, 1959, and Fig. 2).



(*a*)

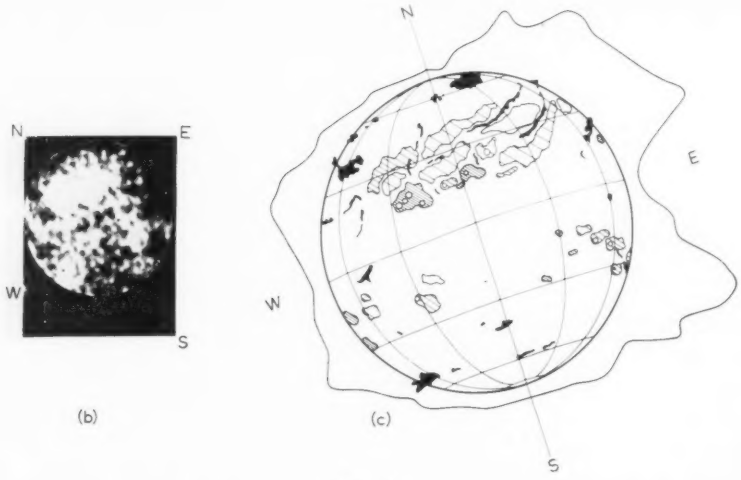


FIG. 2.

(*a*) A weak bipolar facular field with the dark lane and filament, observed at Meudon, 1922. The spectroheliogram was taken in K light (slit-width 0·14 Å), and shows that the structure of the facular field does not differ *essentially* from that of the quiet chromosphere; the main difference is the greater number of bright mottles. *Scale:* 1 mm = 6700 kilometres. South is at the top and West on the left-hand side. (Courtesy L. d'AZAMBUJA, Meudon.)

(*b*) The Sun, photographed in Lyα-light on 1956 May 8, by means of a rocket ascending to 145 km. The photograph was obtained with the Sun-following camera, equipped with LiF optics, through a filter with a pass-band of 40 Å. The original solar image had a diameter of 2·8 mm. (Cf. MERCURE, MILLER, RENSE and STUART, 1956). (Courtesy W. A. RENSE, Boulder.)

(*c*) Daily map of the Sun for 1956 May 8, issued by the Meudon Observatory. Sunspots are shown by small circles, filaments and prominences by black ribbon-like structures. The distance from the irregular line around the Sun to the Sun's limb is proportional to the coronal brightness, measured in 6374 Å at 3 ft distance from the limb. Facular fields are indicated by shaded areas. These fields, seen in K (Ca^+) and Hα, seem to be smaller than those viewed by Ly α.

Table 3

Structure	Average Diameter	Average Life Time	Comments	Comparable Characteristic Length
fine mottles	10^3 km	5 min	spicules seen from above	scale-height in upper part of convection zone; photospheric granules
coarse mottles	$0{\cdot}5$–1×10^4 km	one day?	clusters of fine mottles	thickness of chromosphere
network	5×10^4 km	days	composed of coarse mottles	thickness of convection zone

Thus, these three kinds of elements form a hierarchic system. The main point is, as shown in Fig. 2a, that the principal difference between the normal and the disturbed chromosphere lies in the fact that the latter contains a greater number of bright elements, hence more fine mottles or spicules. ATHAY and THOMAS (1957) also found that the chromospheric facular fields contain a greater number of hot chromospheric elements than the normal chromosphere. The same kind of structures which occur in the quiet chromosphere are visible also in the facular fields.

This leads to a number of suggestions:

(a) the heating of the chromosphere takes place mainly in the fine mottles and in the larger structures composed of them;

(b) since the heating seems to be related to an increased magnetic field, the fine mottles might be the seat of this field;

(c) if this is the case, then the chromospheric network will prove to be identical with a magnetic-field pattern on the Sun; LEIGHTON (1959) confirmed this.

(d) then the true fields of these spicules will be greater than the field measured by BABCOCK for the BM regions; the latter are average fields; the fine mottles, which cover perhaps 5–10 per cent of the solar surface might have a field at least ten times greater, say, 10–50 gauss (see Fig. 7).

The extension of the chromospheric faculae. That the disturbed regions on the Sun have a much larger extent than is usually judged from Hα or Ca^+(K) spectroheliograms, is clearly demonstrated from Figs. 2 (b) and 2 (c), by comparing (*b*), a Ly α picture of the Sun (MERCURE *et al.*, 1956), with (*c*), which is a drawing of the same day of the phenomena visible in the optical region. A greater part of the disc is covered with Ly α faculae than with K faculae, and this shows that the mechanism responsible for the heating of the chromosphere is still effective outside what are usually considered to be the facular limits. A similar conclusion is drawn from the high-resolution pictures obtained by PURCELL, PACKER and TOUSEY, 1960.

If it is assumed, for the moment, that the borders of the visible Lyman α faculae correspond to the line where the radiation intensity in the centre of Ly α exceeds that of the quiet disc by 50 per cent, we find, applying Planck's law, that at that point the radiation temperature in Ly α exceeds that of the quiet Sun by 150°. T_{rad} for the quiet Sun is 6800° and is equal to the excitation temperature for the first two hydrogen levels. It is known (DE JAGER, 1956; ATHAY and THOMAS, 1956) that the corresponding electron temperature of the chromospheric gas is much higher,

about 40,000°: the Ly α line is formed in the transition region between the chromosphere and the corona. With the aid of the available non-thermodynamic equilibrium computations it is estimated that the difference of 150° in the excitation temperatures corresponds to a difference of about 15,000° in the electron temperatures. This remarkable result shows that even far from the centres of the CA's the electron temperatures in the transition region between chromosphere and corona are considerably enhanced by a mechanism that is obviously too weak to produce any observable increase in temperature in the lower chromosphere. This observation must certainly have important consequences for the theory of the heating of the upper chromosphere and the low corona.

4. The Origin of the Magnetic Fields and of Solar Variability

A well developed spot is visible for some weeks, a facular field for some months, or even one year or more, but it must be remembered that the magnetic fields producing the observable phenomena are much more persistent. As Gurevich and Lebedinsky (1945), and Cowling (1946a, b), have shown, such a field has an average life-time of the order of centuries; the conductivity in the Sun's interior is so great that matter can move only very slowly with respect to the fields. If we put in Equation (1) $v = 0$, this equation reduces to the diffusion equation

$$\partial \boldsymbol{H}/\partial t = \frac{c^2}{4\pi\sigma} \cdot \nabla^2 \boldsymbol{H},$$

which becomes, after introduction of the quantities τ and λ:

$$H/\tau = c^2 H/4\pi\sigma\lambda^2$$

With $\sigma = 10^{13}$ e.g.s.; $\lambda = 10^9$ cm, τ becomes 300 years. Thus the field diffuses only slowly; those fields which occasionally become visible must have existed or will exist for several centuries, perhaps in deeper parts of the Sun. It seems quite reasonable to assume that these fields originate in the solar convection zone, which we know extends down to a depth of about 10^5 km. Below the convection zone solar matter is very viscous and moves only slowly.

The problem of how these magnetic fields originate in the convection zone is not yet completely solved. It is often thought that the field originates by conversion of turbulent energy of the solar plasma into magnetic energy by a process of drawing out the lines of force ("spaghetti process"). This process may be active in the convection zone, where it may lead to fields whose strength may be computed by equating the magnetic and the kinetic energy; a computation made for the conditions prevailing in the middle part of the convection zone yields fields of the order of 10^3 gauss, that is, of the same order as the fields of sunspots.

The suggestions made above are perhaps supported by two observations: the diameters of facular fields in the early part of their activity, and the mesh width of the chromospheric network are of the same order as the depth of the convection ($\approx 10^5$ km); and, secondly, the time for a convection element to rise from the bottom of the granulation zone up to the surface is about 30 days, that is, of the same order as the duration of the "active stage" of the activity centre.

The characteristic time for convection elements to pass through the convection zone can be computed from the theory of this zone, as developed by various authors

(VITENSE, 1953; SCHATZMAN, 1953). The usual formula (see UNSÖLD, 1955; equation 55·28) is:

$$v^3 = \frac{R}{2\mu c_p}\frac{L_0}{P}\pi F_{\text{conv}},$$

where we put $R/2\mu c_p = \frac{1}{5}$; πF_{conv} = total solar flux = $10^{10.8}$ (erg. cm^{-2} sec^{-1}), (assuming that in the main part of the convection zone the radiative transfer is unimportant) and L_0 = scale height = $(\mathrm{d}ln\rho/\mathrm{d}h)^{-1}$. From this expression the time T, required for the transit through the convection zone is computed as:

$$T = \int_{-H}^{0} \frac{\mathrm{d}h}{v}$$

where $-H$ is the lower limit of the zone (10^5 km). The evaluation of the integral yields $T \approx 30$ days. This value is uncertain by a factor 3 on the side of higher values, mainly because of the uncertainty in the lower limit H. Nevertheless it agrees so well with the life-time of the first "active" stage of a centre of activity (when spots and flares occur, and when the faculae are of the brilliant condensed type), that one is tempted to think that the occurrence of a bipolar magnetic field at the Sun's surface is connected with a disturbance throughout the whole convection zone. It rises and subsequently disappears by convective motions.

Once such a field has arrived at the surface, it extends upward into the corona, there it produces a *coronal activity region* which at the top of its development may become a "permanent coronal condensation". The process of the escaping of the field from the photosphere may be associated with the occurrence of photospheric loops, described for the first time by KIEPENHEUER (1960).

5. Coronal Activity Regions and Permanent Coronal Condensations

The coronal activity regions are the coronal extensions of the photospheric and chromospheric activity centres. They have a limited life-time of the same order as that of the facular field. Their life-history is closely related to that of the faculae, and the maximum intensity of the coronal intensity measured in the 5303 Å line lags less than a week behind that of the faculae (ALLEN, 1948).

The structure of these coronal activity regions can be studied by various means:

(*a*) Coronagraphic observations show a brightening of the 5303 Å line (Fe XIV, 255 eV) at the same place where we observe the facular field. The red line (7364 Å; Fe X; 233 eV) is also sometimes bright above these regions, but it is often bright at their sides. The increase in brightness of the 5303 Å line suggests that the line is at the maximum of its ionization in the CA's, which would yield an electron temperature of about 1×10^6 °K. These regions often show many fairly stable ray- and arch-structures, about 150,000 km long, often curved, and radiating away from the CA. They tend to expand gradually in the course of hours. Occasionally more rapid motions occur in them.

(*b*) High-resolution radio-observations with cm and dm waves show enhancements that coincide with facular fields. The spectrum of a typical coronal CA shows a relative maximum in the region of the dm wave lengths (Fig. 3). At 22 cm, these activity centres show brightness-temperatures up to $1{\cdot}5 \times 10^6$ °K, (the common value being

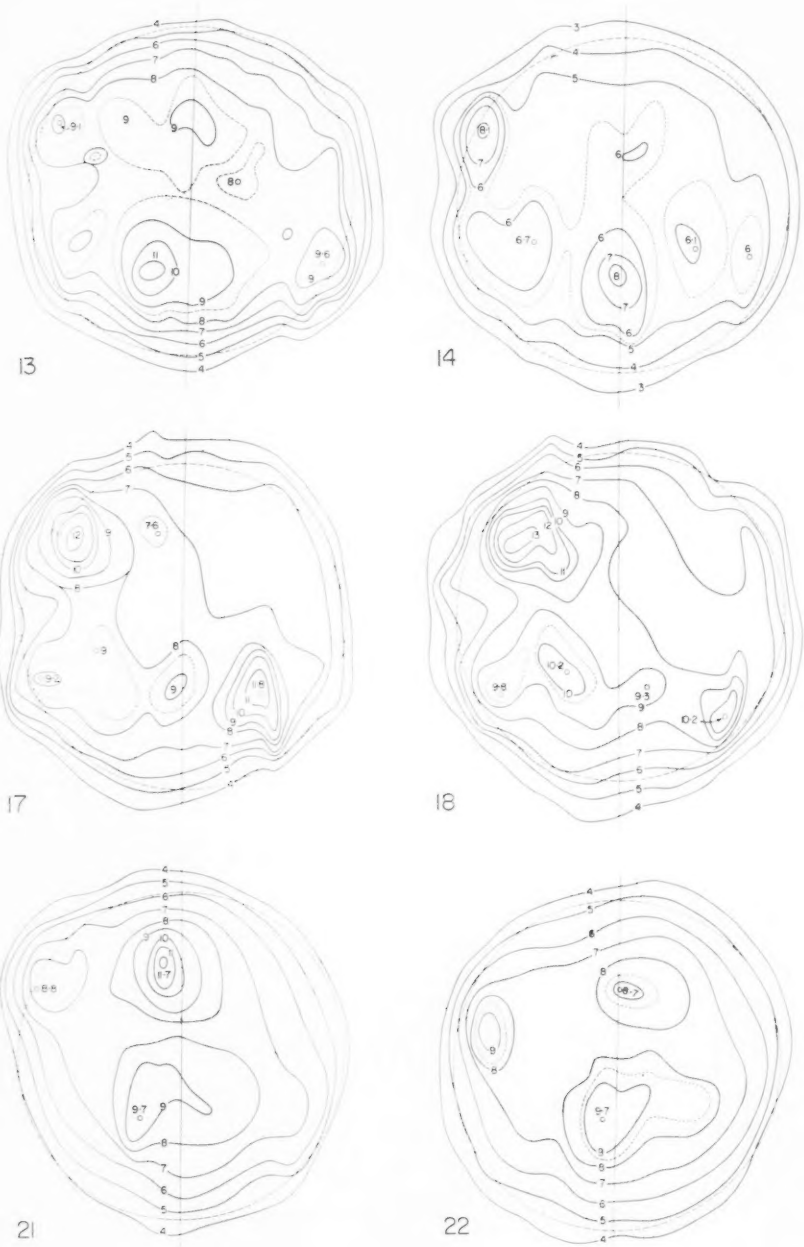


FIG. 3 (*a*)

(*a*) The Radio Sun at $\lambda = 3$ cm. The pictures give the Sun for twelve consecutive days: 1957 July 13–24. North is at the top, East is left. The solid lines connect points of equal aerial temperatures. The dashed line is the limb of the optical Sun. The dates are givenleft below each drawing (July, 1957). (Courtesy V. V. VITKEVICH *et. al.*, Moscow.)

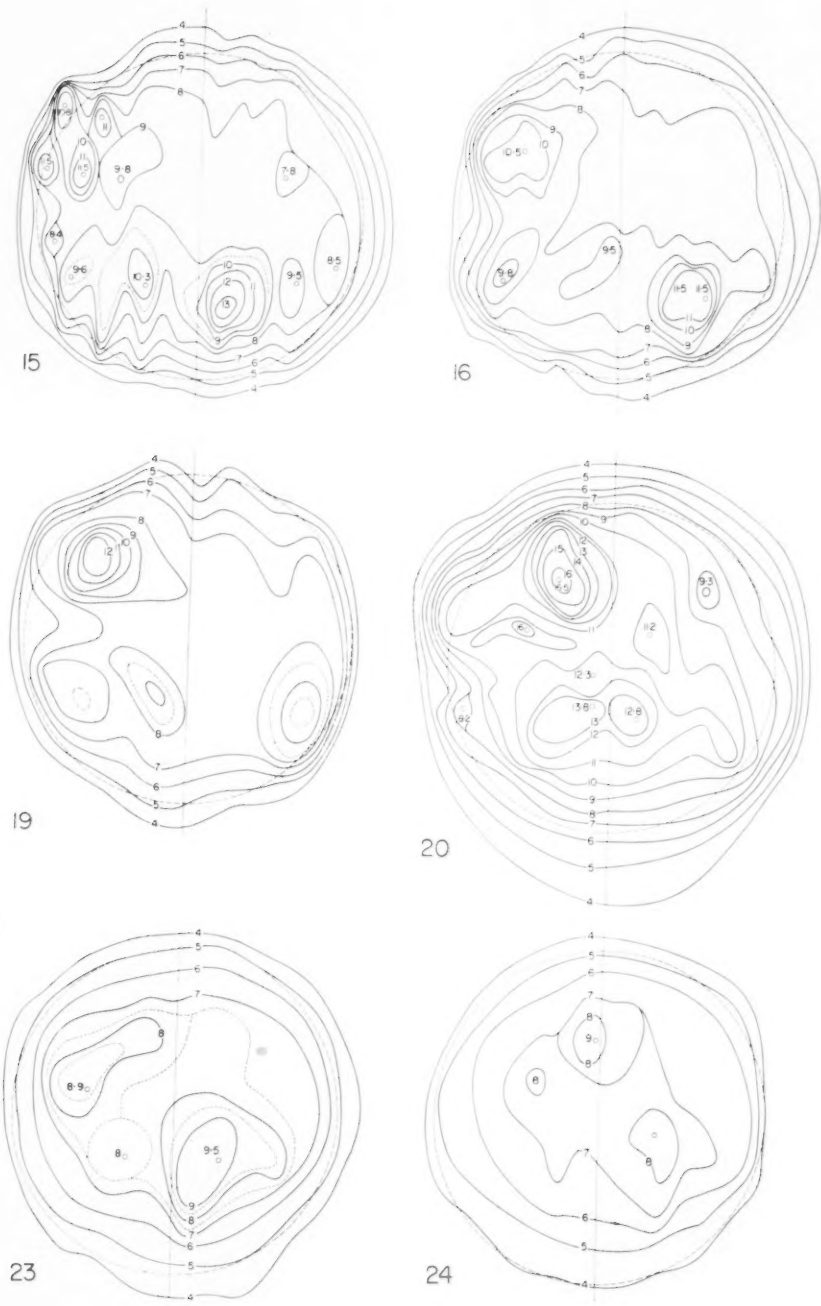


FIG. 3 (*a*) *continued*

6×10^5 °K: CHRISTIANSEN and MATHEWSON, 1959), whereas the brightness-temperature of the undisturbed Sun at 22 cm is 55,000 °K. Such a great increase in brightness cannot be explained by the small increase in temperature suggested by the observations of the green coronal line, but rather suggests an increase of the density.

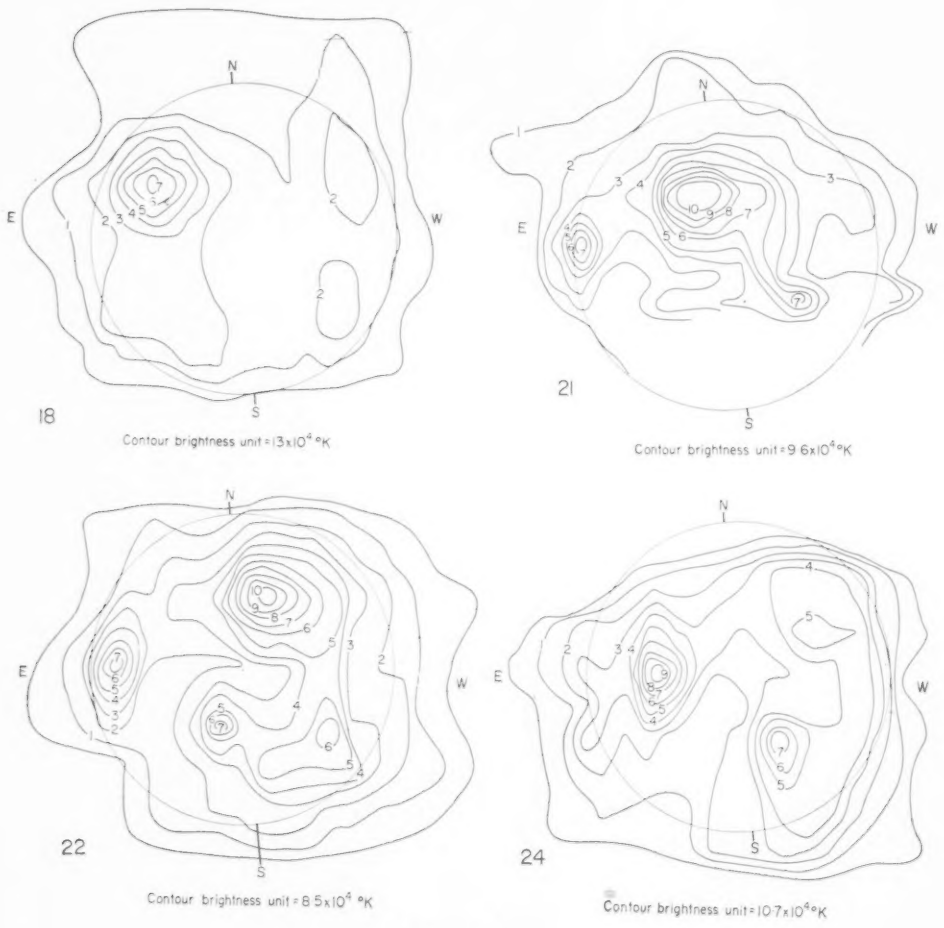


FIG. 3 (*b*)

(*b*) The Radio Sun at $\lambda = 21$ cm. The drawings show the Sun during four days in the period covered by Fig. 3 (*a*), and indicate the good agreement between regions of enhanced intensities at cm and at short dm waves. The dates (July 1957) are given on the left below each drawing. (Courtesy W. N. CHRISTIANSEN, Sydney.)

Roughly, this may be shown as follows: the absorption coefficient of radio-radiation κ_λ is proportional to $N_e^2\, T_{el}^{-3/2}$, whereas the emission of radiation B in the radio-range is proportional to T_{el}, so that the *ergiebigkeit* $\kappa_\lambda B$ is proportional to $N_e^2 T_{el}^{-1/2}$. An increase of T_{el} would thus give a decrease of radiation, unless compensated by an increase of N_e.

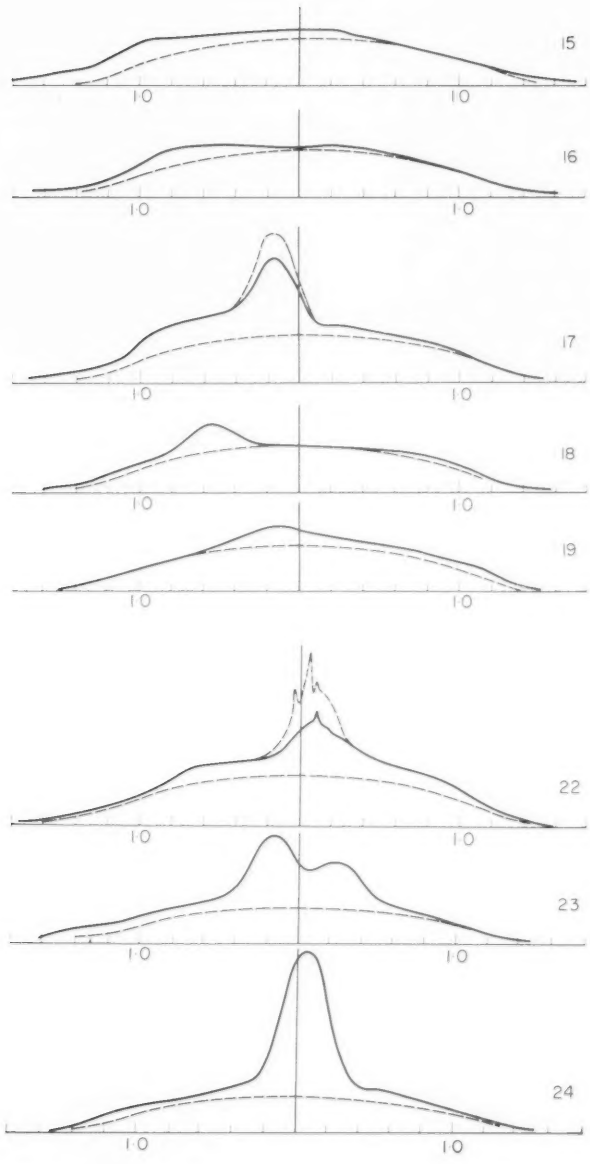


FIG. 3 (*c*)

(*c*) Solar scans at $\lambda = 88$ cm, obtained during the period covered by Figs. 3 (*a*) and (*b*). The scans show that brightness-maxima at short metre-waves do not necessarily correspond to regions of enhanced radiation at shorter waves. East is left, West is right. The scanning angle is 5° in all cases. The dotted line under each curve is the lower envelope of all the scans for 1957. The dates (July 1957) are shown at the right side of each curve. All scans were obtained at about 17 hr 15 min U.T. On July 17 and 22 the Sun's radiation at 88 cm was changing during the period of measurement. Additional dotted scans show details of change which took place in the ten minutes between two successive scans. (Courtesy J. FIROR, Washington.)

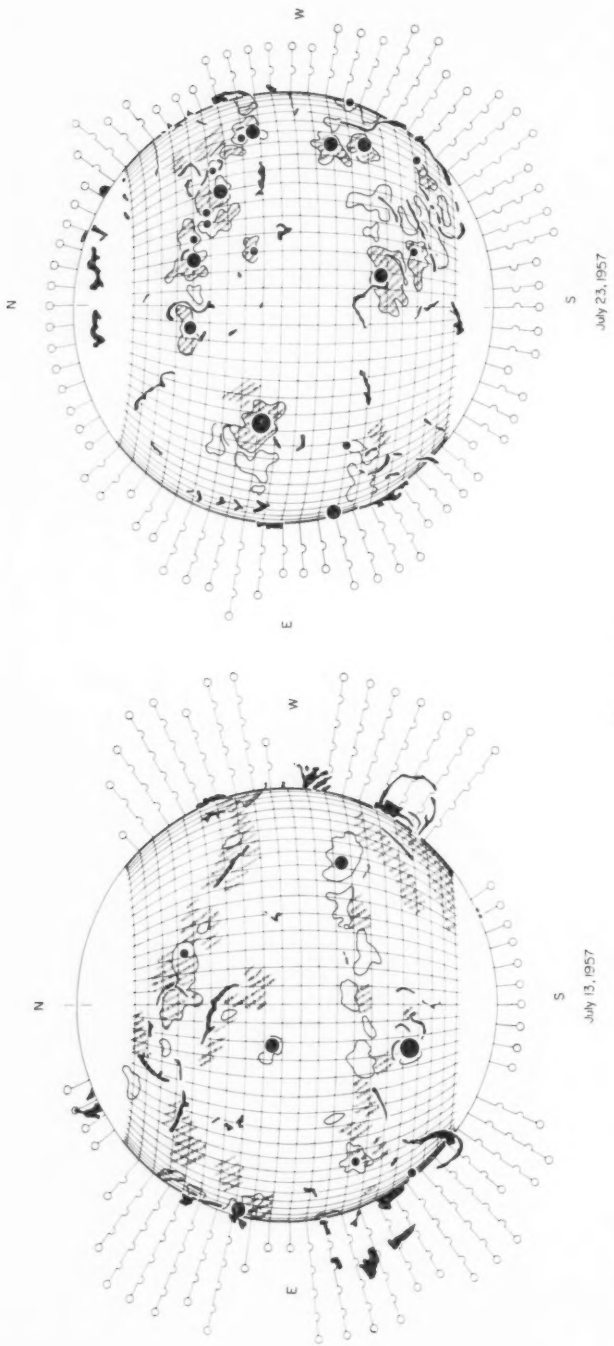


FIG. 3 (*d*)

(*d*) The optical Sun on 1957 July 13, and July 23. (Reproductions from the maps of the Sun, Fraunhofer Institute, Freiburg.)

Application to the case of the 22 cm radiation observations gives the result

$$(N_{e_{CA}}/N_{e_{\text{quiet Sun}}})^2 \lesssim \frac{150{,}000}{55{,}000} = 25$$

(since the influence of T_{el} is virtually negligible); hence $N_{e_{CA}}/N_{e_{\text{quiet Sun}}})^2 \lesssim 5$. For the average CA we have: $N_{e_{CA}}/N_{e_{\text{quiet Sun}}} \approx \sqrt{11} = 3{,}5$.

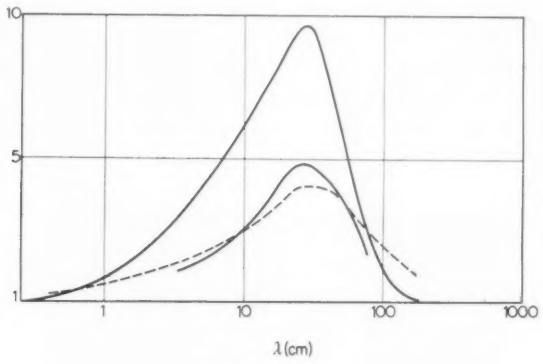


FIG. 3 (*e*)

(*e*) The solid lines give the radio-spectrum of a strong and of an average coronal activity-centre, as compared with that of the undisturbed Sun at the centre of the disc (horizontal line at ordinate 1). The dotted line gives the spectrum computed for an activity centre with the same temperature as the quiet Sun, but with twice its density.

This reasoning is only valid as a first approximation. In a more refined discussion one should take into account the fact that an increase of density throughout the corona causes an increase of the optical depth $\tau_\lambda = \int \kappa_\lambda dh$, and so the centre of the emergent radiation moves upwards to the low corona. Since the radiation of dm waves is emitted in a part of the corona where the temperature increases upwards, this fact causes an increase in the brightness-temperature in the region of enhanced density, and a reduction of κ. The essential fact is that the corona which is transparent for $\lambda = 22$ cm when undisturbed, becomes optically thick for that radiation in CA's. If we assume the corona to be isothermal, then any density-increase at metre waves (which are emitted by the corona) will not produce the slightest increase of brightness temperature. As an example we give in Fig. 3(*e*) the brightness temperature for a coronal activity region assumed to have the same temperature as the quiet corona, but a density greater by a factor two at all heights. This curve suggests that the well-developed coronal activity-regions have a density which is greater than that of the neighbouring quiet corona by a factor of about 4 to 8. The vertical extent of this region of increased density is difficult to determine from radio-observations. At metre waves a *density* increase does not manifest itself in the emergent radiation, if the corona is isothermal. Only T-variations can be discovered.

Thus the coronal activity region presents itself as a region of slightly higher temperature and of considerably greater density (initially 2, increasing to about 6 times that of the normal corona). The projected area of the coronal CA is about equal to that of the faculae. Since an increase in the brightness temperature is only sometimes observed at short wavelengths ($\approx$ 1 m), and for a short time only, the

region of increased *temperature* does not extend much higher than about 10^5 km. The enhanced *density* region *can* extend further, but this must be discovered by other means.

(*c*) The K-coronameter is a newly-designed instrument which records the intensity of the polarized light around the Sun, and which thus enables us to determine the amount of electrons around the Sun and, hence, to find the intensity of the K-corona. With such an instrument, in regular use at the Boulder High Altitude Observatory since 1956, (Newkirk, 1959) it has been found that in the coronal activity-regions the average electron density is about twice that in the quiet regions of the (maximum) corona. The region of enhanced density extends into the corona at least up to one solar radius. At a height of 100,000 km above the photosphere the electron density in the CA is about twice that of the quiet corona, and at 350,000 km and 700,000 km this ratio is 1·8 and 1·5, respectively.

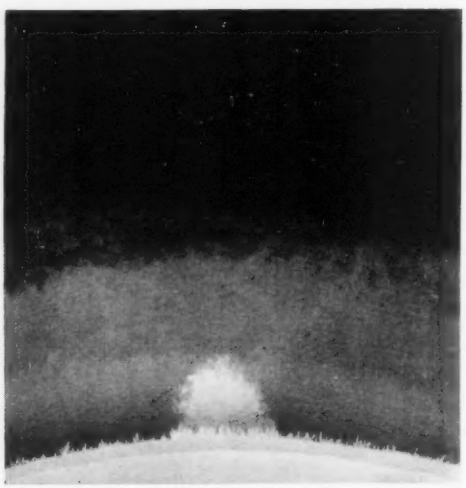


Fig. 4 (*a*)

(*d*) The coronal CA's emit an increased X-ray radiation, as is revealed by the correlation between X-rays (measured with the rocket technique) and the solar coronal activity. This is, furthermore, also shown by the good correlation between the radio-radiation at 10 cm and the (probably X-ray-) radiation which ionizes the ionospheric E-layer (Denisse and Kundu, 1957).

The problem of the X-ray radiation of a coronal activity region, for assumed temperatures of 3×10^6 and 6×10^6 °K, has been considered by Elwert (1958). An increase of the intensity of the radiation has been found below 30 Å; the radiation decreases below about 6 Å. The net result is therefore a hardening of the X-ray spectrum, in agreement with rocket work (Friedman *et al.*, 1958; Chubb *et al.*, 1960).

It is clear from the foregoing, however, that temperatures as high as those here referred to, do not occur in the "normal" coronal activity regions but only in the conspicuous coronal condensations. When spots are at the height of their development, and when flares occur, temperatures up to a few times 10^6 °K may be found, and these may persist for some days; still higher temperatures, characterized by

the yellow line at 5694 Å (Ca XIV, 814 eV), may occasionally be observed in connection with flares. These highly excited regions of the coronal CA are called (after WALDMEIER) the permanent coronal condensations.

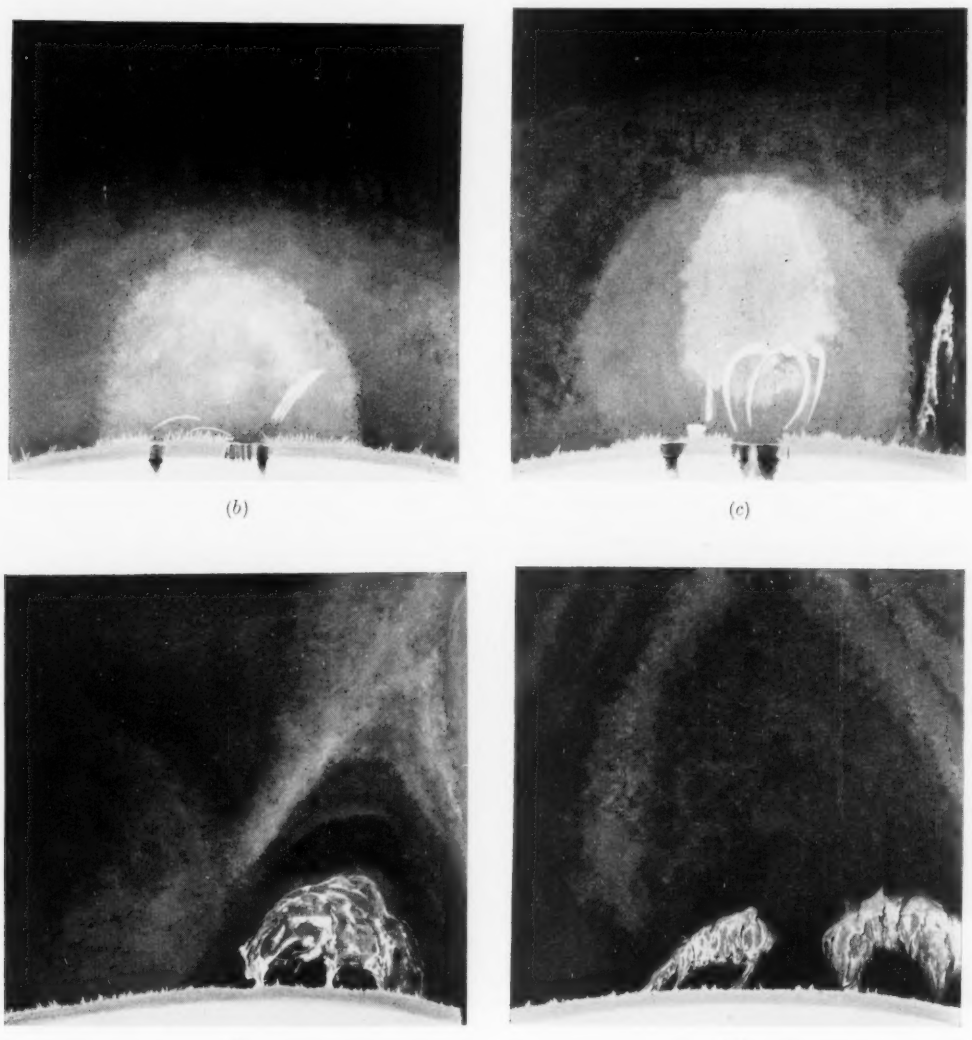


Fig. 4. The development of a coronal activity centre. The five images correspond to those shown in Fig. 1 (*a*); they are meridional sections through the CA. In Figs. 4 (*d*) and (*e*) the centre of activity has been shifted somewhat to the left, in order to show better the coronal rays above the filament.

A *permanent coronal condensation* may live for some days, often in connection with the largest, so-called F-type spot groups. Such condensations show the Ca XV line at 5694 Å (excitation potential 814 eV) and sometimes also a bright continuous spectrum, indicating N_e-values of 0·5 to 2×10^{10} cm^{-3}; they are hemispherical or elliptical. Line profiles indicate temperatures of the order of 2 to 5×10^6 °K. From

radio-observations, at 3 cm, KUNDU (1959) found that the occurrence of flares in a CA is often connected with a radio "hot spot", with a diameter smaller than 75,000 km and with a brightness temperature of 5×10^5 °K (the brightness temperature of the quiet Sun at that wavelength is 15,000°). Probably this, too, is a consequence of the increased density in a permanent coronal condensation.

6. CONDENSATION PHENOMENA IN CORONAL CA's

Three kinds of condensation phenomena are observed in centres of activity: these are the *prominences*, the *sporadic coronal condensations*, and the *flares*. Of the first-mentioned it is known that they are in a kind of equilibrium with their surroundings: at the same horizontal level we have $P_{\text{corona}} = P_{\text{prom}}$, since, roughly, $N_{\text{cor}}\,(T_{el})_{\text{cor}} = N_{\text{prom}}\,(T_{el})_{\text{prom}}$ (ZANSTRA, 1947).

Table 4

	T_{el}	N_e
Corona at 50,000 km	10^6 °K	2×10^8 cm^{-3}
Prominence	15,000 °K	2×10^{10} cm^{-3}

In the vertical direction, prominences are not in hydrostatic equilibrium, since their density is about 10^2 times greater than that of the corona. Therefore, there must be a mechanism that supports the prominences.

It is not yet known whether the sporadic coronal condensations are or are not in hydrostatic equilibrium with their surroundings; and of the flares we know with certainty that they are not. In this Section we shall deal only with the first two kinds of objects; the flares which form a class apart will be discussed in the next Section.

The *prominences* fall, roughly, into two classes of objects:

(*a*) the spot prominences, (short-lived, short and unstable) which occur mostly in the neighbourhood of a spot, hence in the central part of a CA; and

(*b*) the long, stable, and long-lived quiescent prominences, which as a rule are found at the borders of a CA, or divide a CA into two parts of opposite polarity, lying in the "dark lane" between these two parts.

The *quiescent prominences* may represent an equilibrium configuration between the external magnetic fields and the internal kinetic energy at the border of the CA. With the data given in the table above, and with $v_{\text{kin}} \approx v_{\text{turb}} \approx 5$ km/sec, we find $\frac{1}{2}\rho v^2 = 2 \times 10^{-3}$ erg cm^{-3}. Equating this to the magnetic energy density $H^2/8\pi$, one finds $H = 0{\cdot}2$ gauss, which seems a very acceptable value for the border of a CA. The theory for the support of quiescent prominences has been worked out by KIPPENHAHN and SCHLÜTER (1957) who showed that a region of stability occurs in those regions where the magnetic lines of force are horizontal or slightly concave. Equation (4) shows that in such a case a downward motion will be rapidly inhibited; also, that if a prominence is not wholly ionized (as seems to be the case for the quiescent prominences), the motion of fall will be very slow. This is perhaps the reason why these prominences are often met inside a dark zone between the two magnetically opposite parts of a BM region, or at the high-altitude side of a CA. (The combination of the magnetic field of the CA and the general polar field of the Sun might produce horizontal lines of force at the high-latitude side, but not at the

low-latitude side.) Also the fact that most quiescent prominences are slightly inclined *away* from their CA, may perhaps find a magnetic explanation (BECKER, 1956).

Since inside the CA the magnetic energy-density is greater than at the borders, by at least a factor 100, the sunspot prominences and their motions are completely determined by the magnetic field. The magnetic lines of force are "frozen into the matter", but since the magnetic energy-density is much greater than the kinetic energy of matter, it is the matter that has to follow the lines of force—not *vice versa*. Thus the prominences will move mainly when the lines of force move, that is, when the configuration of the sunspots in the spotgroup changes.

The condensation mechanism. The fact that many prominences originate by condensation from the corona is supported by the observation that the corona seems to be more tenuous around quiescent prominences than in undisturbed parts, and also by the fact that many more prominences stream downwards than upwards.

Consider a knot of coronal material. Under normal circumstances, there is equilibrium between the gain of energy from a mechanical flux coming from the chromosphere, and the loss of energy by conduction downwards (cf. DE JAGER, 1959b), as well as by radiation (mainly X-ray radiation at very short wave-lengths) and by evaporation. Suppose that for some reason such a knot decreases in temperature. Then, at temperatures below about 6×10^5 °K the knot will mainly radiate in the form of free-free and free-bound Lyman emission, at the following rates:

$$\text{free-free} \quad \epsilon_{ff} = 1{\cdot}4 \times 10^{-27} T^{\frac{1}{2}} N_e^2 \text{ ergs cm}^{-3} \text{ sec}^{-1} \qquad \ldots . (5)$$

$$\text{free-bound} \quad \epsilon_{fb} = 5{\cdot}4 \times 10^{-22} T^{-\frac{1}{2}} N_e^2 \text{ ergs cm}^{-3} \text{ sec}^{-1} \qquad \ldots . (6)$$

These rates are equal for $T = 3{\cdot}9 \times 10^5$ °K.

If the initial cooling mechanism is such that the temperature decreases to below that value, the knot becomes unstable (KIEPENHEUER, 1951). The radiation given in (6) produces a further cooling of the knot, and this continues to decrease in temperature; but then ϵ_{fb} increases further, and so it goes on. The knot will cool down and will finally reach so low a temperature that it emits lines like Hα and becomes visible to the eye. This process continues until eventually stability is reached again: the emission of radiation is then compensated by an inflow of energy from coronal particles that are absorbed by the prominence. The mean free path of particles of coronal temperatures, computed on the basis of a Coulomb interaction (see UNSÖLD, 1955), with a collisional cross section:

$$Q = (4\pi/9) Z^2 e^4 k^{-2} T^{-\frac{1}{2}}$$

(Z = particle charge), is of the order of 10 cm only in a prominence of $T = 15{,}000$ °K. Thus each coronal particle is stopped in the prominence; its energy $2kT$ is absorbed. Equality between this absorption and the emission of radiation is reached at $T = 15{,}000°$ for a prominence of the thickness of the order 10^4 km, which is equal to the observed thicknesses of prominences.

An alternative mechanism has recently been proposed by KIEPENHEUER (1960), who was able to show that the corona becomes unstable with increasing pressure: if, in a given volume, the pressure increases, the gas temperature decreases, which leads to a further compression, and so on. The initial pressure-increase may be due to the movements of magnetic fields, which carry the "frozen-in" gas with them.

The *sporadic coronal condensations* form a special class of condensation in the corona. They occur in the CA, as a rule at the top of its development, often in

the "permanent coronal condensations". At that time one often observes loop- and knot-type prominences which occur some hours or days after flares. They indicate the time of greatest activity of the CA. A peculiarity of the sporadic condensations is that they can often be observed simultaneously in the coronal lines 5303 Å and 6374 Å *and* in the Hα-line. In the first type of picture they are more diffuse than in Hα. The situation seems to be the following: the hot condensation exhibits a temperature- and density-gradient reaching from the corona to the Hα-emitting part of the condensation, and passing through the part that emits coronal lines. Data on the density and temperature variation in the sporadic condensations are not yet known (see also Section 8).

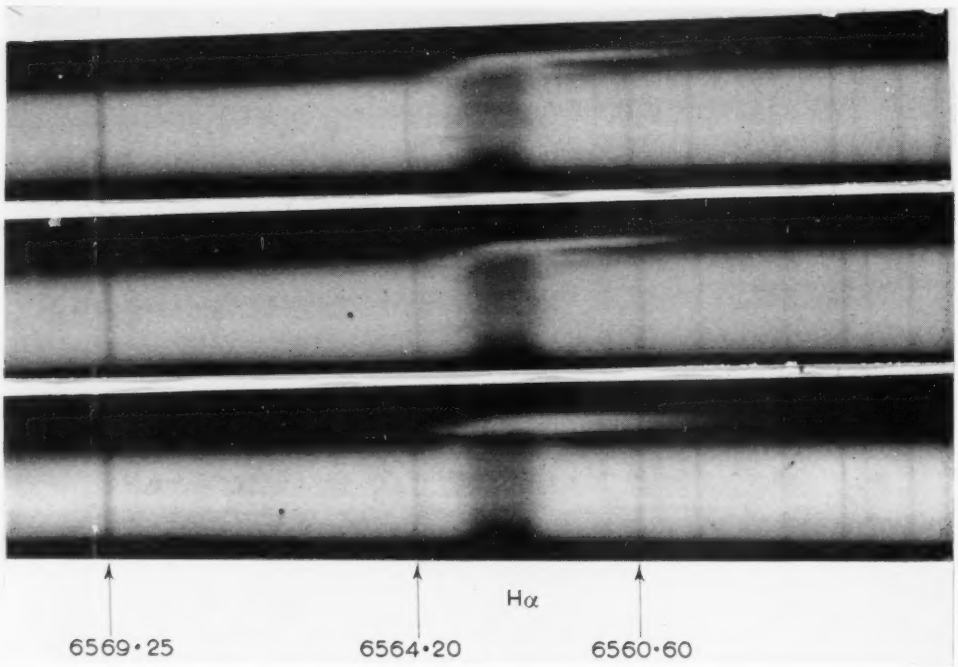


Fig. 5

Limb Hα-flare of 1958 June 3, detached from the chromosphere and showing considerable Doppler motion. The flare has a weak continuous spectrum (invisible in the reproduction) and a limb moustache on the first two exposures. Time of exposures: 13 hr 53 min 00 sec, 53 min 30 sec and 55 min 10 sec. Vertical scale: 1 mm = 7000 km.

7. Flares and Associated Dynamic Phenomena

Flares form one of the most striking phenomena in centres of activity. They are sudden and short-lived brightenings of a small region of the disc, and are mainly observable in light of the cores of some strong Fraunhofer lines. That they are hardly or not at all visible in white light indicates that flares are not photospheric phenomena. The question might be raised whether they are chromospheric phenomena; but on the basis of limb observations of flares this does not seem to be very likely. Flares, as a rule, occur in the chromosphere or the lower part of the corona: the average height might be of the order of 7000–15,000 km above the photosphere.

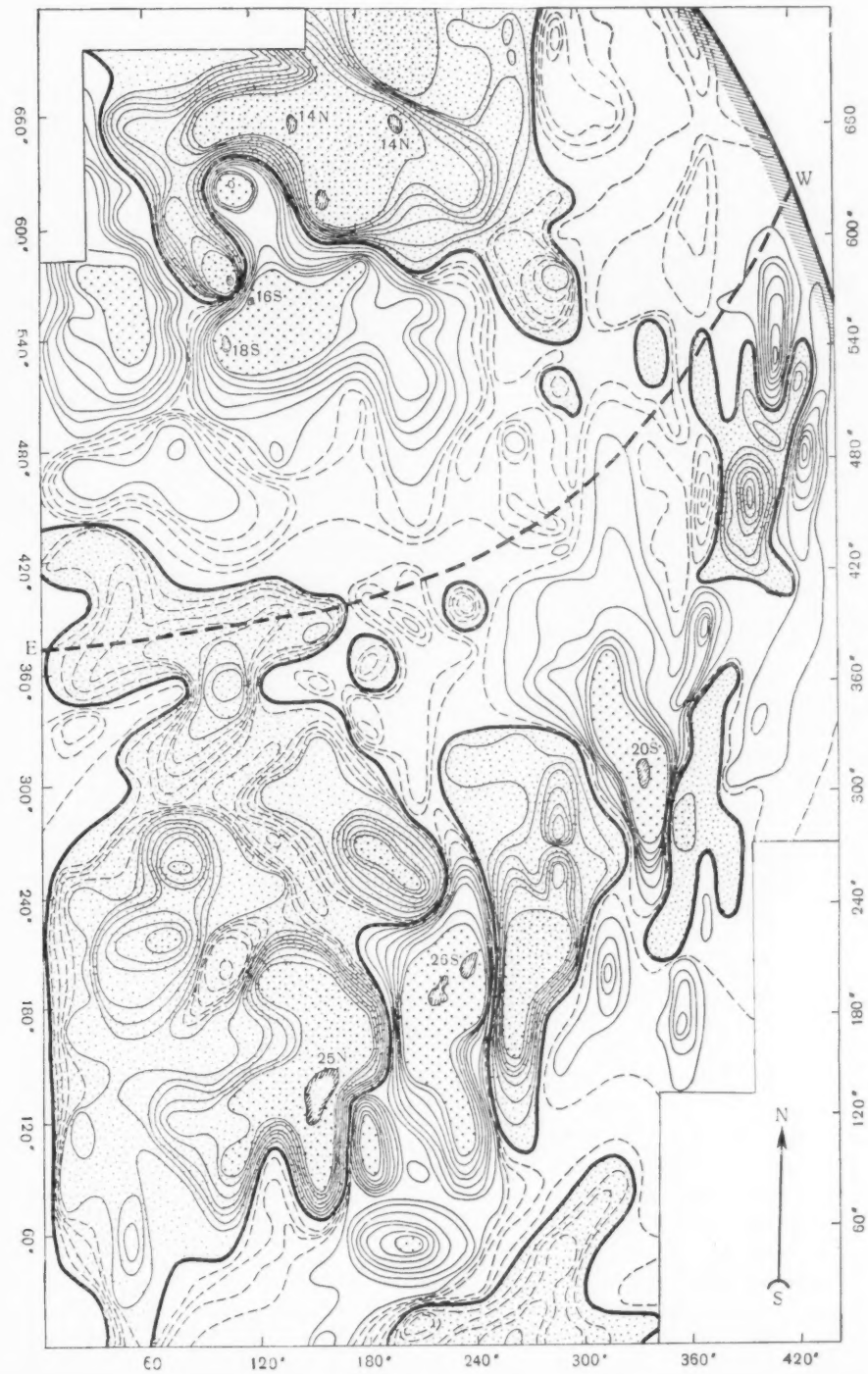


FIG. 6 (*a*)

(*a*) Detailed magnetogram of a solar activity region near the Sun's limb (1957 September 14); two consecutive lines differ by 5 gauss; V. E. STEPANOV, 1958.

Also, the fact that many limb flares can easily be observed *detached* from the chromosphere shows that at least part of the flares are low coronal phenomena (Fig. 5). The temperatures and densities are $T_{el} \approx 10{,}000$ °K, and $\log N_e \approx 13$ to 13·5 (densities valid for the flare knots and only observable with fairly high resolving power). These data make it clear that the chief characteristic of a flare is its great density (not, for instance, its temperature). The density is greater than that of the corona by a factor 10^4 to 10^5, and greater than that of the chromosphere by a factor 10^3 to 10^4.

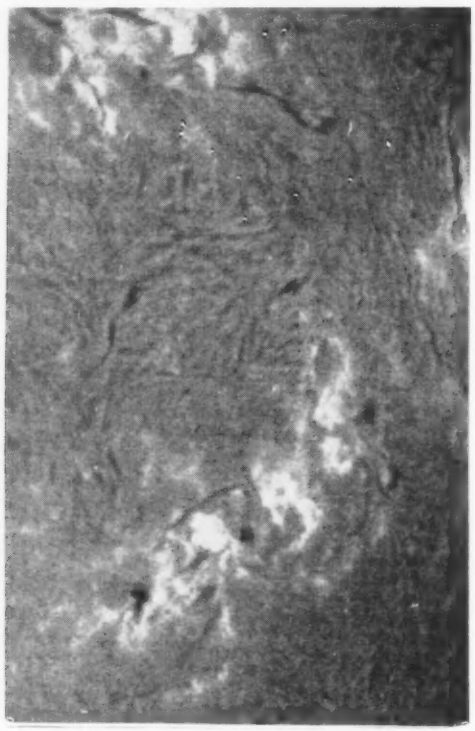


FIG. 6 (*b*)

(*b*) The Sun's disc in the same region, as shown in Fig. 6 (*a*), photographed through an Hα-filter. (Courtesy A. B. SEVERNY.)

Since the temperature of a flare is smaller than that of the surrounding corona by a factor 10^2, the pressure at the same level is greater by a factor 10^2 in the flare than around it. Thus, a flare is dynamically unstable; by diffusion the pressure equality tends to re-establish itself. This explains its short life-time. That the temperatures of flares are also lower than those of the corona might be caused by a cooling-mechanism similar to that considered for the prominences. This mechanism has not yet been investigated in detail.

According to Equation (6), the emission of coronal material is proportional to $T^{-\frac{1}{2}}N_e^2$. Consider a knot of coronal material suddenly attaining a greater density by some mechanism, and hence a greater N_e-value. Because of the greater number of ion-electron collisions, the emission from the coronal knot increases, and hence the

temperature decreases. Not every increase of density is sufficient to "break" the coronal stability, but as soon as the density increases above a certain lower limit, the increase of radiation sufficies to make the knot thermally unstable. The temperature decreases further. In analogy to the case of the prominences this decrease of temperature progresses until stability is reached again.

The next point is the question how the *great density* to which the cooling is due is reached. This increase of density is perhaps caused by magnetic compression. The clue to the solution of the problem has been given by observations of SEVERNY (1958a, b) at the Crimean Observatory, who showed that the origin of flares is connected with an instability of the magnetic field. Observation of the Sun's magnetic field with a high image-resolving power (Fig. 7), before and after a flare, have shown that the

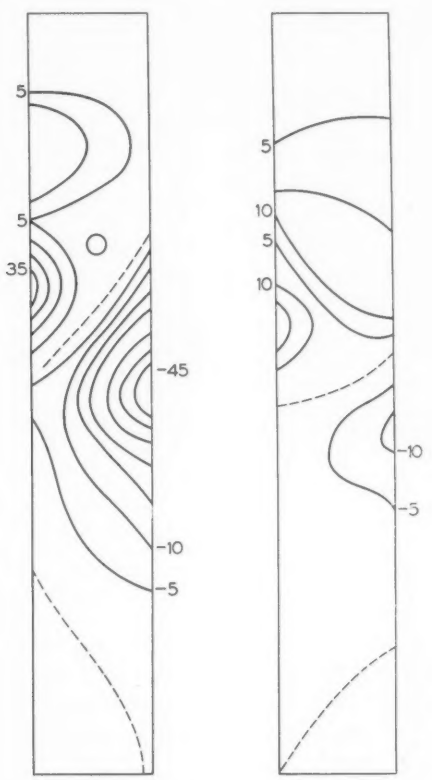


FIG. 7 (*a*)

(*a*) Two magnetograms of the same disc region before and after a flare. Left: at the time of the flare (small circle), 1957 August 27; *Right*: one hour later.

magnetic fields may differ before and after the flare (Fig. 8). It seems that the flare is connected with a collapse of the field. Since in chromospheric-coronal plasma the lines of forces are completely "frozen" into the matter, the matter collapses with the field. The fact that flares are certain to be caused by changes of the field has already been suggested by GIOVANELLI (1938), who discussed the correlation between changes of the areas of sunspots and the occurrence of flares, and also by BRUZEK (1959), who found indications that splitting of a spot into several smaller

ones, or especially into two of opposite polarity, favours the production of flares. Differing from these results is an observation by HOWARD, CROGG, and BABCOCK (1959), who did not record any change during the appearance of a 3+ flare. The problem was also studied by BUMBA (1958). According to SEVERNY, the density increases rapidly during the compression of the field; it can be shown that the internal gas pressure increases less rapidly than the magnetic pressure, thus leading to instability and to the field's collapse. This continues till the linear compression has increased to about 0·01 (leading to a density increase of 10^2 or 10^6, in the cases of linear or isotropic compression, respectively). The knots of high density thus formed

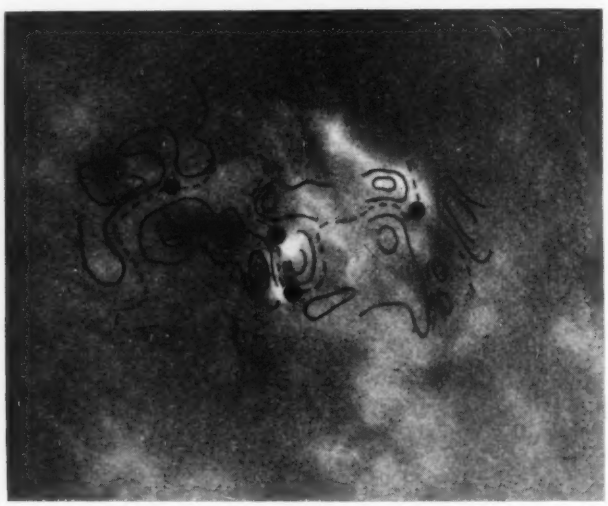


FIG. 7 (*b*)

(*b*) Relation between chromospheric structures and the magnetic field. 1957 August 30, 09h 28m. Dots indicate the positions of brightness maxima; these occur mainly in neutral regions. The lines connect points of equal field strength components; two consecutive lines differ by 10 gauss. Dotted: the field strength zero-lines.

may be identified with the so-called moustaches, observed as "hydrogen bombs" by ELLERMAN (1917) and extensively studied by SEVERNY and others (Figs. 5 and 6). The name "moustaches" was initially given to the spectra of tiny structures (diameter $\approx 1''$, or less) producing an Hα-emission line at photospheric or low chromospheric level: the emission line is only visible in the form of two emission *wings* at the side of the otherwise undisturbed Hα-absorption line (we know that the core of the Hα-line, up to $\Delta\lambda \approx 0{\cdot}5$ Å, is formed in chromospheric layers above about 2000 km). Later this name was often applied to other tiny emission-structures, and also to those which are emitting at levels above about 5000 km (where the radial optical depth in the Hα-centre is about unity). These latter structures produce a narrow emission line without absorption core.

An essential property of the moustaches seems to be that their profiles seem to be Dopplerian (SEVERNY, 1958) and that they suggest expanding motions of 500–1000 km/sec. This should perhaps be interpreted as follows: the initial compression of the chromospheric matter, which causes the formation of chromospheric granules, eventually leads to an expansion at the moment when the kinetic energy density

becomes larger than the magnetic energy-density, so that from then onwards the lines of force, still frozen into the matter, move with it. This expansion seems to be related to the formation of shock waves at the moment when the state of higher contraction is reached. Many of the dynamic phenomena that originate after the development of a flare might be due to this expanding motion.

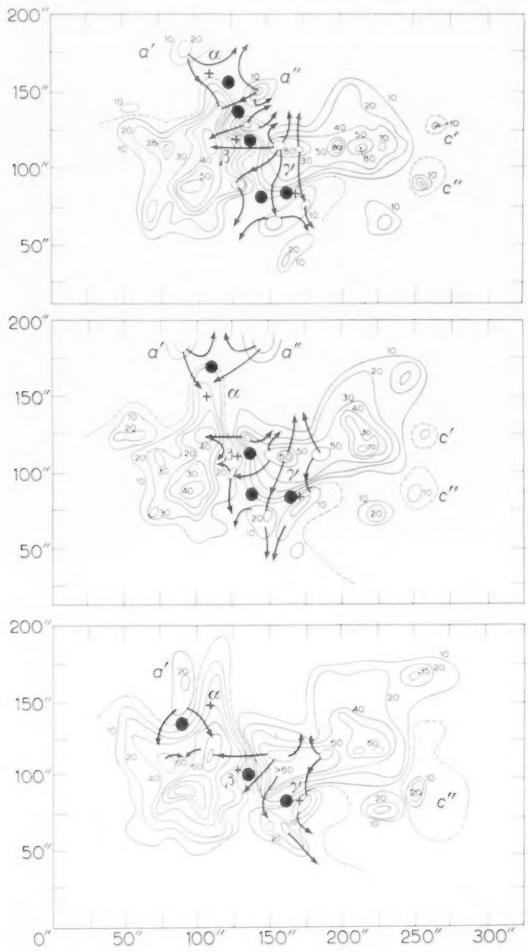


FIG. 7 (*c*)

(*c*) Three magnetograms of a disturbed region, obtained at the Crimean Observatory, 1958 August 22.

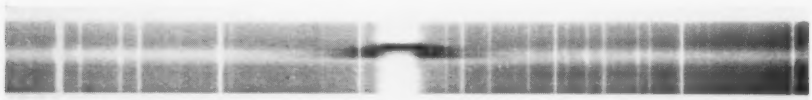
(1) 14 hr 50 min local time, before the appearance of a flare of importance 2 which started at 16 hr 18 min;

(2) at 17 hr 20 min after the flare;

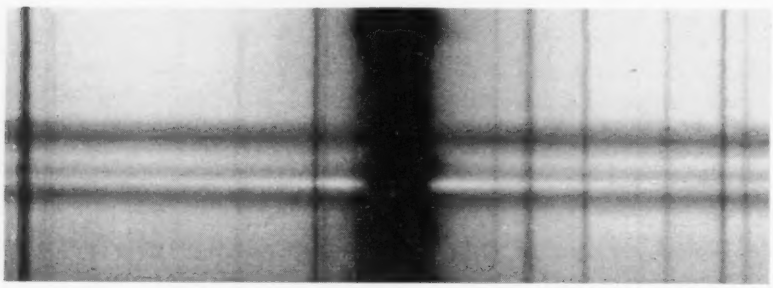
(3) at 18 hr 00 min after the onset (at 17 hr 56 min) of a flare of importance 3.

The dotted lines connect the points where the line-of-sight component of the magnetic field is zero. The dots are approximate positions of neutral points, while the crosses (+) are the positions of the main bright centres of the flares. The arrows approximately represent the field tubes. The solid lines connect points of equal magnetic field strength and differ by 10 gauss. Each map records an area of 300 × 200 seconds of arc; the recording takes about 30 min. (Courtesy A. B. SEVERNY, 1958a, and 1959.)

The occurrence of *high velocity motions* during and after flares represents one of the important problems of solar physics. The observations in Hα show that flares proper, as a rule, present only slow expanding motions, with velocities of the order of 10 or several tens of km/sec. The flare itself sometimes seems to expand more rapidly, but this is rather an expansion of the flare-excitation region than a real mass motion. Sometimes, however, the flares themselves also show motions up to 100 km/sec and even more; (see Fig. 5).



(*a*)



(*b*)

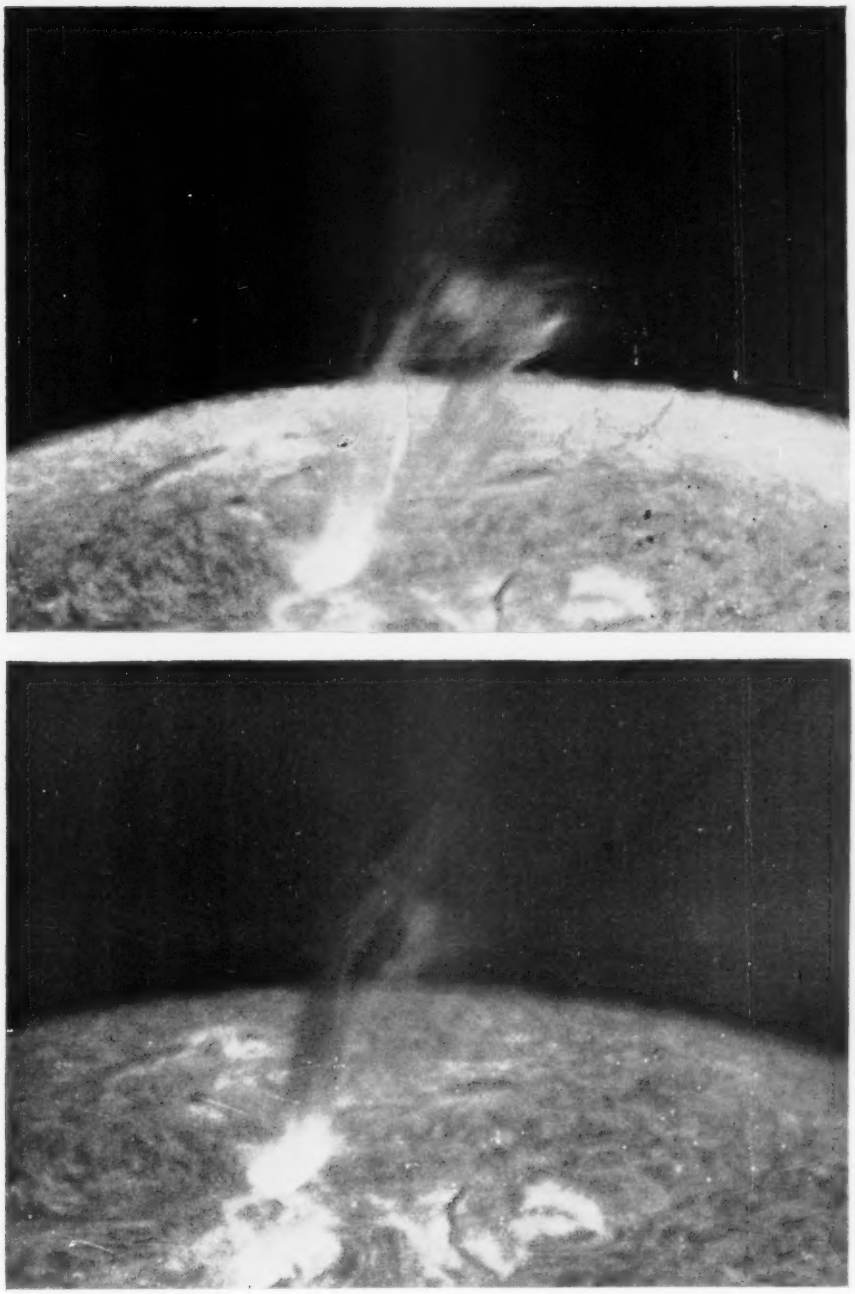
FIG. 8. Moustaches in disc spectra of flares.
(*a*) 1957 July 16 (Courtesy A. B. SEVERNY, Crimean Observatory).
(*b*) Spectrum obtained at the Anacapri station of the Fraunhofer Institute. (Courtesy K. O. KIEPENHEUER.)

In the second stage of the flare's development (mostly after the "flash-stage" which corresponds to the greatest contraction of the flare granules) one often observes flare *surges*, revealing up and downward motions with speeds up to several hundred km/sec. These speeds are too small to lead to an ejection of matter. What one often observes is a kind of pulsating motion, in which the upward motion, followed by a downward one, repeats itself various times.

Speeds, so great that they can lead to a real ejection of matter, requiring a speed close to the photosphere exceeding 600 km/sec, are observed only in their first stage.

Although the optical evidence for velocities so great that they lead to an ejection of matter is rather rare, the *radio observations* tell us that such velocities are rather common. Out of 308 flares, mostly micro-flares, observed between November, 1955 and July, 1956. about one fourth were accompanied by Type II events (LOUGHHEAD, ROBERTS and MCCABE, 1957). Apart from the Type II and Type III radio events, related to flares, there are some other radio-phenomena that may occur together with flares. They are presented schematically in Fig. 10 and in Table 5.

It should be borne in mind that the complete group of radio-events correlated with flares, as shown in Fig. 10, occurs only very rarely in connection with the *same* flare. Groups of Type III bursts occur fairly often; Type II bursts are less frequent; noise-storms (consisting of Type I bursts) occur in perhaps one-third of the cases. Type IV



(*a*)

(*b*)

FIG. 9 (*c*) and (*d*). Surges of 1957 April 18, photographed in the centre of Hα at the Anacapri station of the Fraunhofer Institute.

(*a*) 13 hr 25 min. (*b*) 13 hr 29 min.

(*c*)

(*d*)

FIG. 9 (*c*) and (*d*). Surges of 1957 April 18, photographed in the centre of Hα at the Anacapri station of the Fraunhofer Institute.

(*c*) 13 hr 40 min. (*d*) 13 hr 42 min in the red wing of Hα (Courtesy K. O. KIEPENHEUER.)

bursts were observed about once a month at Nançay (France) during the 1957 maximum of spot-activity, the highest maximum ever observed. The Type V bursts are quite common: they follow (about one in every three) groups of Type III bursts. The occurrence of all five types of bursts in connection with the same flare has never been observed, mainly because Type IV and Type I bursts occur seldom together.

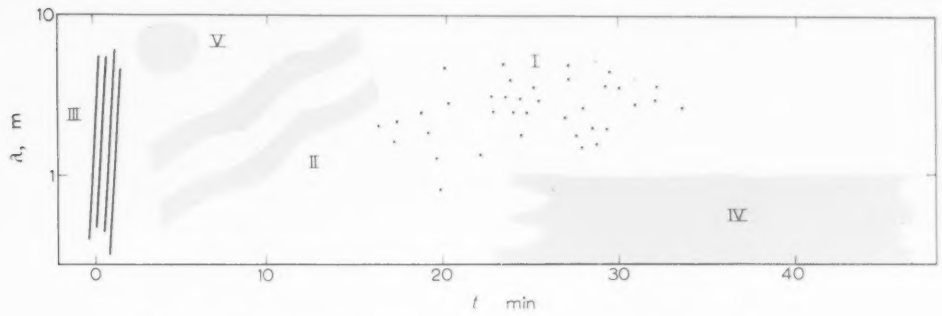


FIG. 10. Radio-events which may occur in connexion with solar flares.

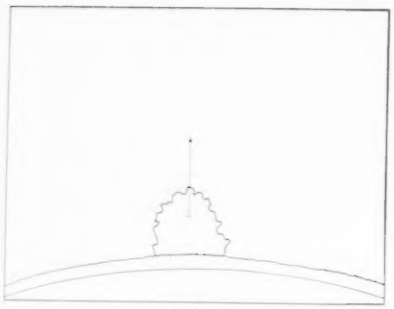
The importance of Type III and Type II bursts for the understanding of the events taking place in a CA at the same time as a flare, lies in the fact that they clearly indicate the presence of high-velocity motions which thus seem to occur much more often than would be concluded on the basis of optical evidence alone. The velocities of the particles can be inferred from the observed frequency shift, at least in principle, by assuming that the emitted frequency is the plasma frequency:

$$f_0 = 8{\cdot}93 \times 10^{-3} \sqrt{(N_e)} \text{ Mc/sec},$$

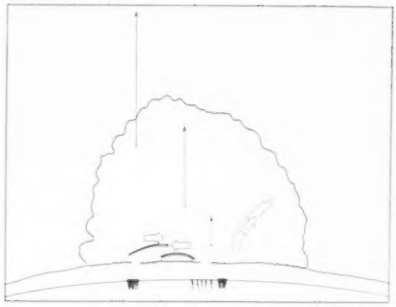
and further that the density-height distribution in the activity centre is the same as that of the quiet Sun. In this way initial velocities were found for the Type II and Type III bursts, respectively, between 500 and 10^4 km/sec. WILD, SHERIDAN and TRENT (1958) succeeded interferometrically in observing the motions that produce the bursts, and found that the velocities obtained were greater than those deduced formerly, namely, of the order of $1{\cdot}5 \times 10^3$ km/sec and 10^5 km/sec, respectively. The optical velocities of the phenomena associated with Type II bursts rarely exceed 500 km (GIOVANELLI and MCCABE, 1958). Hence, the velocities of the Type II events are similar to those observed for the solar particles that produce the large geomagnetic storms.

The velocities observed for the Type III bursts which previously were assumed to lie between $c/10$ and $c/3$, and for which we now take values between $c/3$ and c, are still unconnected with those of the cosmic ray bursts: the latter are due to particles, moving with virtually the velocity of light.

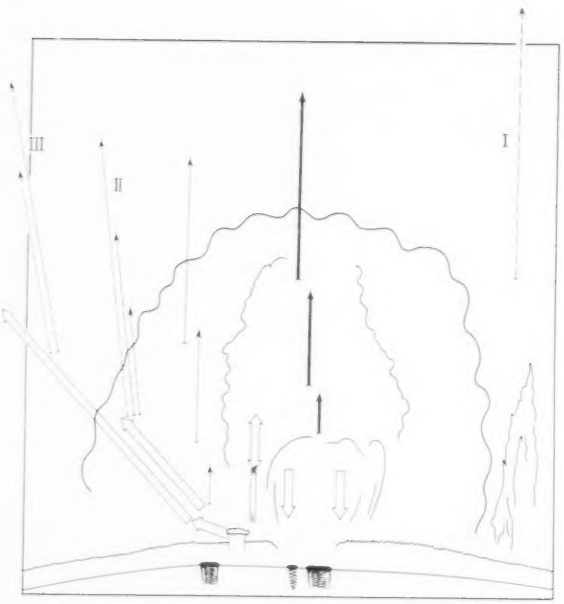
Another conclusion to be drawn from the comparison of the new, directly measured velocities with those deduced formerly on the basis of the quiet corona density-law is, that the density-height distribution in a CA differs from that in the quiet Sun by a factor of at least two—a result also found qualitatively from K-coronameter observations (NEWKIRK, 1959), and from the observation of coronal activity regions at radio and X-ray wavelengths.



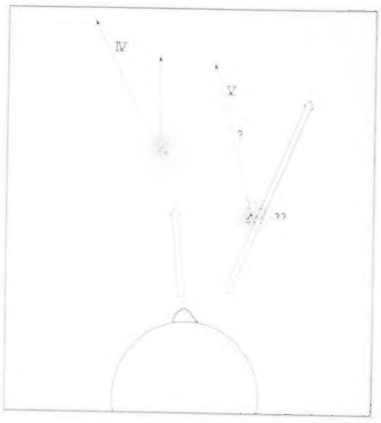
(*a*)



(*b*)



(*c*) (1)



(*c*) (2)

(*d*)

(*e*)

FIG. 11 (*a*)—(*e*)
Radio emission and mass-motions in a Centre of Activity.
For detailed description see caption to Fig. 11 (f) on p. 175.

The problem of the acceleration of these particles has not yet been solved; various valuable investigations have been published which, however, do not lead to a prediction of the actually observed large velocities. One of the main difficulties in all theories is the induction law: in the highly conductive plasma, with which we are concerned, the effect of induced currents is to oppose the changes to which they are due.

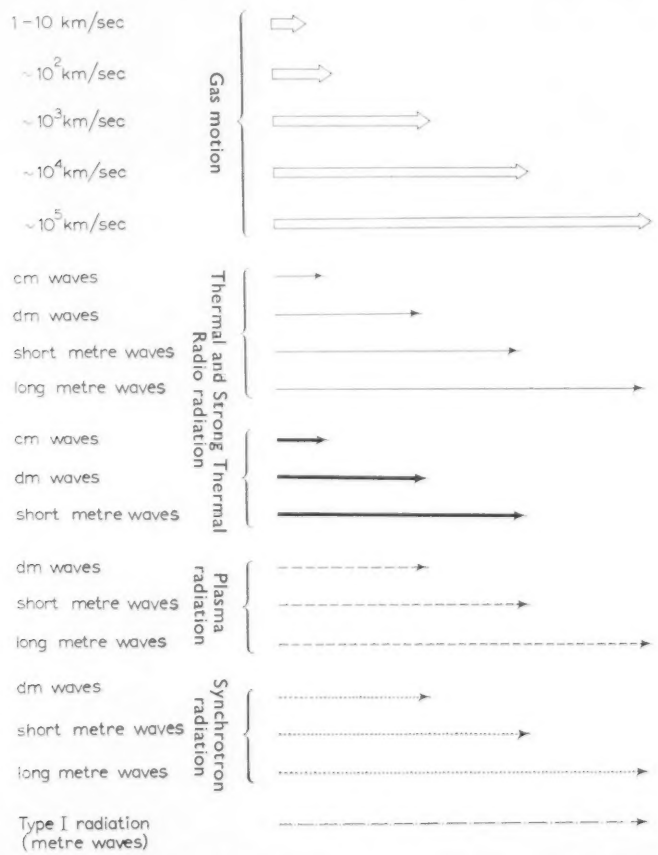


FIG. 11 (*f*): *Explanation of the six Figures* 11 (*a*)–(*e*) *on p.* 174. Radio emission and mass motions in a Centre of Activity:—The figures correspond to the sets of five figures shown in Fig. 1 (*a*) as well as in Figs. 4 (*a*)—(*e*), and give schematically the regions of origin, the wavelength, the emission-mechanism, of the various observed radiations, and the velocities of the mass motions. Figure 11 (*c*)(2) [which is on a smaller scale than Fig. 11 (*c*) (1)], shows the Type IV region. Single arrows: radio-radiation; the wavelength is indicated by the length, the emission-mechanism by the kind of arrow. The bases of the arrows indicate the region of origin. Double arrows: mass motion; the lengths give the order of magnitude of the velocity. In Figures 11 (*a*)–(*e*) the numerical lengths of these various types of arrows differ somewhat from the symbolic diagrams above; but the relative arrow-sizes, within each of the Figures (*a*)–(*e*), indicate the actual situation.

8. "HOT SPOTS" IN THE UPPER CORONA

Recent observations of the corona by means of radio observations and with the K-coronameter have confirmed the previous radio results, that even at its maxmum activity a coronal condensation does not present temperatures much in excess of about 2×10^6 °K, but that the density ρ is appreciably greater than that of the

Table 5

Radio-events often correlated with flares

Type	Other name	Duration	Time (with reference to flare)	Frequency Range	Band-width (Mc/sec)	Speed (through corona)	Polarization	Probably emitted by
III		some sec	at start	metre and dm spectrum	some tens	10^5 km/sec	partly	plasma radiation
II	outburst	some min	after flash	meter and dm spectrum $\lambda \geqslant 50$ cm	some tens	10^3 km/sec	unpolarized	plasma radiation
I	stormburst	$\leqslant$ 1 sec	end-phase	mainly metre spectrum	3–7		unpolarized	?
IV		1 hr	end-phase	dm and short metre spectrum $\lambda \geqslant 50$ cm	200	10^3?	polarized	synchrotron radiation
V		some min	after Type III bursts	long metre waves	25	3×10^3 ?		?

surrounding quiet corona. Higher temperatures occur only occasionally in smaller regions, accompanied at that time by emissions of the 5694 Å line (Ca XV, $\chi_i = 816$ eV), the line of highest excitation potential in the solar spectrum. The occurrence of this line is correlated with the presence of flares and of the "sporadic condensations", highly active prominences of the knot-, loop- or funnel-type. These prominences tend to show a very lively condensation-tendency and intensive downward streaming. Matter condenses in knots at some height in the corona and from these knots there is a continuous streaming down, sometimes along straight paths, but often also along curved ones. There is a continuous replenishment from the corona, supplying the matter that descends down to the chromosphere. This continuous streaming from the corona (through the loop- or knot-prominences) to the chromosphere may last for some tens of minutes, but it may also continue for some hours. These prominences and the so-called yellow coronal regions (emitting the 4695 Å line) are strongly related to flare activity; according to WALDMEIER, they may originate after the occurrence of a flare and reach maximum-intensity half an hour afterwards. The decrease of intensity is slower than that of the flare.

Therefore, following WALDMEIER, we have to distinguish between

(*a*) the "permanent coronal condensation"—highly excited regions in the corona, with a life-time of some days and with kinetic temperatures of the order of 4×10^6 °K; and

(*b*) the "sporadic condensations" with loop- or knot-structure. They occur within the permanent ones, live shorter and have lower temperatures (see Sections 5 and 6).

The local brightenings described above, that may occur for a shorter time within the permanent condensations, have still high temperatures; they are real "hot spots" in the corona, with life-times of the order of an hour.

There are *other indications* of the occurrence of *energetic regions* in the upper corona shortly after the occurrence of flares. The precise relationships between the previously described phenomena and those to be dealt with below are still to be investigated.

(*a*) The Type IV radiobursts show a continuous spectrum that slowly increases in brightness after a flare, to reach maximum after ten or several tens of minutes. It may last for some hours. The spectrum is sharply limited on the side of low frequencies. Interferometric observations made mainly at Nançay (France) show that the Type IV bursts often occur in the corona after Type II motions. This motion moves an excited region up to a great distance from the Sun, 0·5 to 4 $R_\odot$, and at that distance a coronal region is then formed that is able to emit the continuous radiation described above. These regions obviously occur in rather tenuous parts of the corona.

Any theory on the Type IV bursts has to explain how a coronal region at that height above the photosphere can emit a continuous radiation which occasionally can exceed that of the whole quiet Sun by a factor of the order 100; it has also to explain why the spectrum has a cut-off at low frequencies, and why this radiation can persist for one hour or so. It is clear that thermal radiation can be left out of consideration. Because of the local low density of the corona, plasma oscillations of that frequency do not seem very probable either. There is a much greater probability that we are dealing here with synchrotron radiation, as suggested by BOISCHOT and DENISSE (1958).

A discussion summarizing the theory of the synchrotron radiation was given by TAKAKURA (1959). The synchrotron radiation is produced by electrons moving in helical orbits under the influence of the magnetic field. Each electron, in uniform circular motion with frequency ω, emits energy with frequencies ω and at the higher harmonics. One observes a smooth spectrum because of the spread in the velocities. The reduction of the radiation at high frequencies is due to the strong resonance absorption near the gyro-frequency; the "cross-over" frequency (average $\approx$ 600 Mc/sec) has, therefore, perhaps to be identified with the gyro-frequency. This, in turn, determines the magnetic field-strength to about 200 gauss. For such a field, the life-time of the synchrotron electrons is determined by radiation damping (loss of energy by radiation) for low particle-densities, and by collisional damping for high densities. Assuming a density of the order of 10^8 cm^{-3}, the observed life-time turns out to be comparable with the assumed field for $\beta = v/c = 0{\cdot}25 - 0{\cdot}90$. For smaller values of β the influence of collisions reduces the life-time.

Another approach to the problem is to identify the "cut-off" at about 600 Mc/sec with the characteristic "maximum frequency" of the synchrotron radiation spectrum (BOISCHOT and DENISSE, 1958; WILD, 1959). In this case the necessary field-strength becomes much smaller, of the order of a few gauss only.

Thus the Type IV radiation can be emitted by a region in the medium corona with a magnetic field of some hundreds of gauss. The problem how a strong field can reach such high coronal regions may be solved by recalling that Type IV bursts are often preceded by Type II bursts, hence by upward-moving streams of particles. It has been suggested that the magnetic field in high coronal regions is brought there by this particle stream.

(*b*) CHUBB, FRIEDMAN, KREPLIN and KUPPERIAN (1957) have found that during flares the solar X-ray radiation is considerably hardened, and that during strong flares the short wavelength-limit of X-ray radiation goes down to a wavelength between 1 and 2 Å. The emission of an intensive radiation at wavelengths down to 1 or 2 Å has been predicted by ELWERT (1956); more detailed calculations were given by ELWERT (1958). This radiation is produced by the excitation of K electrons of the very abundant element Fe. In order to emit radiation of that wavelength, the coronal regions must have electron temperatures up to about 10×10^6 °K. Satellite observations will have to inform us whether this radiation also shows a time lag with respect to the optical flare.

(*c*) Finally, reference is made to the cosmic radiation, emitted by the Sun after great flares. The five great cosmic radiation bursts that have been observed up to now all started after the beginning of a flare, with time-lags between $0^h.2$ (February 23, 1956) and $2^h.5$ (July 25, 1946). Since cosmic rays particle of the relevant energies have virtually the velocity of light, the cosmic rays were emitted by the Sun $0^h.2$ to $2^h.5$ after the flare.

These various kinds of observations of the enhanced solar radiation, both at radio-wavelengths (Type II bursts), in the optical range (sporadic condensations), on X-ray, and in the range of cosmic rays, form a challenging set of observations, waiting for a detailed observational and theoretical attack.

It may be useful to summarize the various kinds of coronal structures met in this review and to add to the summary some tentative values on temperature and density.

Table 6

	T (million degrees)	log N_e (cm^{-3})	H gauss	Life-Time	referred to in Section
Quiet Corona	1–2	9 5	< 2	long; 11-years' cycle in coronal density	
Coronal Activity Region	2–2·5	2 . . .50 × quiet corona	10	4 months	2; 5
Coronal Condensation	≈4	10–20 × quiet corona	?	some days	5
"Hot spot" in coronal condensation	5–10	?	?	20^{min}–2^{hrs}	8
"Hot spot" Type IV event	?	8?	200?	20^{min}–2^{hrs}	8
"Hot spot" Type V event	?	?	?	some minutes	7

9. Later Phase of a CA

The "active phase" of a CA ends when the spots have disappeared. The further life-history of a CA is of a much less complicated nature, and is roughly characterized by an increase in the size of the facular area and of the bipolar (BM) region, followed after one solar revolution by a gradual decrease of the facular brightness and of the field, and later by a disintegration of the faculae. The field loses its bipolar character and it may in rare cases even become a unipolar magnetic (UM) field, while still increasing in area, but decreasing in field strength. In general, the UM phase is not reached; the BM field weakens until it becomes indistinguishable within the surrounding complex pattern of weak fields.

The long quiescent filaments reach their greatest lengths at these stages; they occur as a rule on the high-latitude side of the BM and are gradually pushed upwards perhaps by the general expansion of the field itself. Just as with the field, the filaments orientate themselves more and more parallel to the Sun's equator, due to the Sun's differential rotation. "Sudden disappearances" of the filaments may often occur, under the influence of forces originating from neighbouring flares of new CA's. In the first case the disappearances are rarely final; often the filament re-appears after one or two days in the same form and position. This shows that the magnetic field conditions, that give rise to the filaments, remain virtually unchanged during the disturbance.

A special problem in this respect is offered by the *coronal rays*. These rays are often observed during eclipses at mid-phases of the solar-activity cycle, and may extend to many solar diameters. These rays seem to be correlated with the quiescent filaments, hence with the CA's in their last phases.

A frequently discussed problem concerns the question whether the coronal rays represent an outstreaming motion. This problem has not yet been solved. The suggestion that the central disc-passage of a ray is followed by a geomagnetic disturbance, after a few days, is not confirmed (Piggot and Allen, 1957); and never have appreciable Doppler motions been seen in ray-spectra, although these should be observable since not all the rays should be strictly perpendicular to the line of sight.

However, since a ray has a greater density than its surroundings, and since it has a finite life-time, it is in any case certain that a ray is related to material motion. Its form, broad at the basis and squeezed at higher levels, suggests an electromagnetic explanation; the "pinching" might then be related to an increase of the velocity.

In order to obtain some insight into this problem, it is perhaps best to consider the most elementary form of coronal rays: the polar plumes. Their forms strongly suggest a relationship with the polar solar magnetic field and in fact, their orientation has often been used to study this field. They have a diameter at the basis of about 7000 km, and an average life-time of 15 hr (WALDMEIER, 1955). Their density exceeds that of the surrounding corona by a factor of the order 5 to 9, and the density gradient is similar to that of the corona (DZUBENKO, 1957). It has been suggested (VAN DE HULST, 1950) that the polar rays have their origin in a density-excess at the basis of the corona. This density-excess is prevented from sideward diffusion by the magnetic field: the particles spiral along the field-lines, and they are driven upwards by the divergent field-lines. The problem is now to find out where the density-excess comes from. The answer may be indicated by recalling that life-times and diameters of the polar plumes closely resemble those of the coarse mottles of the chromosphere. However, the density-excess above a normal chromospheric coarse mottle seems to be less than the value observed in the polar rays. Also the number of rays is appreciably smaller than that of the coarse mottles, so that only a few, and perhaps only the most pronounced of the polar coarse mottles seem to produce rays. Perhaps, the rays are related to the so-called polar faculae, which have similar *life-times* and *sizes* as the rays and mottles, but show a much greater brightness than a normal mottle.

10. The Mass-Balance of the Corona and of the Coronal Activity Centres

It is not the least surprising that solar CA's are related to the emission of particles. The very existence of the solar corona is the result of an equilibrium between the inflow of mechanical energy to the corona and the outflow of energy by (1) backward conduction to the chromosphere, (2) X-ray radiation, (3) evaporation of the hot particles and (4) condensation of matter into prominences and downfall to the chromosphere. It is further known that the inflow of mechanical heat from a facular field is greater than that from the quiet solar regions (see Section 3). Since a local increase of the kinetic temperature of the corona will in any case produce an increase in the number of particles in the high-velocity tail of the Maxwellian velocity distribution curve, the amount of matter that evaporates increases also. It does not seem improbable, therefore, that any CA may be related to emission of particles. The problem arises why this seems to apply more to old CA's and why young CA's produce rather short coronal structures, often arch-like, but never long rays. This may be due to the magnetic field; in a bipolar field like that of all young CA's, where the magnetic lines of force are of horseshoe shape and partly parallel to the surface, the radial emission of particles is for a time prevented, since sideward diffusion is prohibited by the magnetic field.

This picture is supported by the proposal made by H. D. and H. W. BABCOCK (1955) that the so-called "M regions", (hypothetical regions, made responsible for the emission of particles leading to recurrent geomagnetic storms (BARTELS, 1934)), may provisionally be identified with the UM regions. An example of such a correlation can be found in the literature (BABCOCK, 1955). However, in this connection we

should immediately add that MUSTEL (1957, 1959) recently advanced arguments that recurrent geomagnetic storms are produced by facular fields, hence by CA's in the first part of their life, without, however, excluding the possibility of particle-emission at the time when the facular emission has vanished. (See also MUSTEL and TSAP, 1958.)

Finally, a remark concerning the energy balance and the origin of the corona. The above discussion suggests that loss of matter occurs mainly in the CA's, because these have the highest temperatures: a small increase in temperature causes a great increase in the loss of matter, since the coronal temperature is close to the limiting value for which the corona can still be stable ($1{\cdot}5 \times 10^6$ °K, according to PIKELNER, 1951a, b, c). See also PIKELNER (1958). Notwithstanding this continuous loss of matter, the corona CA's have a greater density than the quiet corona: there must be a source which makes good the loss and even helps to produce a density excess.

The whole life-history of the coronal CA's makes it evident that this unknown source must be the lower chromosphere or photosphere, not something in the interplanetary space, since a coronal CA starts its life in the lowest parts of the corona. The next point is by what kind of mechanism the matter is fed into the corona. The only aspects by which the chromospheric CA's differ from the quiet chromosphere are the magnetic field and the greater brightness (the latter caused by a greater number of fine mottles, as high-resolution spectroheliograms show).

It thus seems that the mechanism by which matter is injected into the corona is connected with the fine mottles, hence with the spicules of the chromosphere; these occur in greater numbers in the faculae than outside. The fact that a coronal CA has about 2–5 times the density of the quiet corona agrees with the fact that the spicular number-density in a facular field is also about 2–5 times the number density in the quiet chromosphere. That the coronal density is greater at sunspot maximum than at minimum also indicates that near the maximum coronal-activity a greater amount of matter is sent into the corona than near minimum. That this matter comes principally from below is also shown by the time-lag between the moment of maximum-intensity of the coronal features in an activity centre, and that of the corresponding photospheric phenomena.

That the excess density of a coronal activity centre is lost in a time of the order of some months shows that the loss of matter must take place at a fairly rapid rate, perhaps in a characteristic time short compared to some months. This matter is lost either by *evaporation*, or by *condensation* into prominences, and is being sucked into the chromosphere. The time-constant can be determined fairly precisely. BIERMANN, 1941 (see also 1956, 1958, 1959), from comet observations, estimated that the corona gets lost in a time of the order of one day. PARKER (1960) described theoretically this "solar wind" and confirmed that on the average the particles remain in the corona no longer than a few days.

Summarizing, we should like to stress the fundamental importance of the magnetic field in all phenomena in a centre of activity. The very origin of a CA is caused by the appearance of the field at the surface of the Sun; the faculae and the spots all represent this field in different grades of intensity. The flares are produced by sudden changes in the central parts of the bipolar magnetic field; the prominences are condensations, perhaps initiated by field changes, and are prevented from falling by the magnetic field. Also the permanent and the sporadic coronal condensations, and perhaps also the long coronal rays, find their explanation in the magnetic field connected with the CA's.

The CA disappears with the field, not because the field has diffused—the solar magnetic fields are very persistent—but because the field simply disappears, presumably downward into the deeper parts of the Sun.*

REFERENCES

Allen, C. W.	1948	*Terr. Magn. atmos. Elect.*, **53**, 433.
Athay, R. G. and Thomas, R. N.	1956	*Astrophys. J.*, **124**, 586.
	1957	*Astrophys. J.*, **125**, 804.
Babcock, H. D. and H. W.	1955	*Astrophys. J.*, **121**, 349.
Babcock, H. W.	1957	*Astrophys. J.*, **126**, 224.
Bartels, J.	1934	*Terr. Magn. atmos. Elect.*, **39**, 281.
Becker, U.	1956	*Z. Astrophys.*, **40**, 65.
Beggs, D. W. and Klüber, H. von	1956	*Nature, Lond.*, **178**, 1412.
Biermann, L.	1941	*Vjschr. Astr. Ges.*, Leipzig, **76**, 194.
	1956	*Proc. Symposium on Electromagnetic Phenomena in Cosmical Physics*; I.A.U. Symposium No. **6**, p. 248, 1956.
	1958	*Rev. Mod. Phys.*, **30**, 975.
	1959	Astrophys. Symposium Liège 1959 on *Stellar Evolution* (Discussion). Mém. in 8°, vol. **17**, No. 409, p. 439.
Böhm, K.-H.	1958	*Z. Astrophys.* **46**, 219.
Boischot, A. and Denisse, J. F.	1958	*C.R. Acad. Sci. Paris*, **245**, 2194.
Bruzek, A.	1959	*Z. Astrophys.*, **47**, 191.
Bumba, V.	1958	*Izv. Krymsk. Astr. Obs.*; (*Crimean Astrophys. Obs.*), **19**, 105.
Christiansen, W. N. and Mathewson, D. S.	1959	In *Paris Symposium on Radio Astronomy*, 1958, ed. R. N. Bracewell, I.A.U. Symposium No. 9, p. 108; Stanford Univ. Press.
Chubb, T. A., Friedman, H., Kreplin, R. W. and Kupperian, J. E.	1957	*J. geophys. Res.*, **62**, 389.
Chubb, T. A., Friedman, H. and Kreplin, R. W.	1960	*Space Research*: COSPAR Colloquium at Nice; p. 695; North-Holland Publishing Co., Amsterdam.
Cowling, T. G.	1933	*Mon. Not. R. astr. Soc.*, **93**, 90.
	1946a	*Mon. Not. R. astr. Soc.*, **106**, 218.
	1946b	*Mon. Not. R. astr. Soc.*, **106**, 446.
	1957	*Magnetohydrodynamics*, Interscience Tracts on Astronomy; p. 65; New York.
de Jager, C.	1956	In *Solar Eclipses and the Ionosphere*, ed. W. J. G. Beynon and G. M. Brown, p. 174, Pergamon Press, London.
	1957	*Bull. Astr. Insts. Netherlds.*, **13**, 133.
	1959a	In *Handb. der Phys.*, **52**, pp. 80–362; Springer, Göttingen.
	1959b	In *Paris Symposium on Radio Astronomy*, 1958, ed. R. N. Bracewell, I.A.U. Symposium No. 9, p. 89, Stanford Univ. Press.
Denisse, J. F. and Kundu, M. R.	1957	*C. R. Acad. Sci. Paris*, **244**, 45.
Dzubenko, N. I.	1957	*Astr. Zh.*, **34**, 379.
Ellerman, F.	1917	*Astrophys. J.*, **46**, 298.
Elwert, G.	1956	*Z. Astrophys.*, **41**, 67.
	1958	*Z. Astrophys.*, **44**, 112.
Friedman, H., Chubb, T. A., Kupperian, J. E., Kreplin, R. W. and Byram, E. T.	1958	*Astrophys. J.*, **128**, 738.
Giovanelli, R. G.	1938	*Astrophys. J.*, **88**, 204; see also *Nature, Lond.*, **141**, 746.
Giovanelli, R. G. and McCabe, M. K.	1958	*Austral. J. Phys.*, **11**, 130.
Gurevich, L. and Lebedinskii, A.	1945	*C. R. Acad. Sci. U.S.S.R.* (NS), **49**, 92.
	1946	*Fiyichesk. Zh.*, **10**, 327 and **10**, 425.
Howard, R., Crogg, Th., and Babcock, H. W.	1959	*Nature, Lond.*, **184**, 351.

Acknowledgement. My sincere thanks are due to Mr. J. Damen Sterck for the drawings.

KIEPENHEUER, K. O.	1951	*Publ. Astr. Soc. Pacific*, **63**, 161.
	1953	*Convegno di Scienze Fisische Matematicehe e Naturali Sep.* 1952, *Fondazione Alessandro Volta, Atti dei Convegni*, **11**, 105.
	1960a	*Z. Astrophys.*, **48**, 290.
	1960b	*Z. Astrophys.*, **49**, 73.
KIPPENHAHN, R. and SCHLÜTER, A.	1957	*Z. Astrophys.*, **43**, 36.
KUNDU, M. R.	1959	In *Paris Symposium of Radio Astronomy*, 1958, ed. R. N. BRACEWELL; I.A.U. Symposium No. 9, p. 229; Stanford Univ. Press.
LEIGHTON, R. B.	1959	*Astrophys. J.*, **130**, 366.
LOUGHHEAD, R. E., ROBERTS, J. A. and MCCABE, M. K.	1957	*Austral. J. Phys.*, **10**, 483.
MAKITA, M.	1958	*Publ. Astr. Soc. Japan*, **10**, 1.
MERCURE, S., MILLER, S. C., RENSE, W. A. and STUART, F.	1956	*J. geophys. Res.*, **61**, 571.
MUSTEL, E. R.	1957	*Astr. Zh.*, **34**, 120.
	1959	*Astr. Zh.*, **36**, 222.
MUSTEL, E. R. and TSAP, T. T.	1958	*Izv. Krymsk. Astr. Obs.* (*Crimean Astrophys. Obs.*), **18**, 74. (See also **16**, 67, 1956).
NEWKIRK, G.	1959	In *Paris Symposium on Radio Astronomy*, 1958; ed. R. N. BRACEWELL, I.A.U. Symposium No. 9, p. 149; Stanford Univ. Press.
PARKER, E. N.	1960	*Astrophys. J.*, **132**, 821.
PIDDINGTON, J. H.	1956	*Observatory*, **76**, 21.
	1957	*Mon. Not. R. astr. Soc.*, **116**, 314.
PIGGOT, W. R. and ALLEN, C. W.	1957	Ninth Report on Solar-Terrestrial Relations.
PIKELNER, S. B.	1951a	*Izv. Krymsk. Astr. Obs.* (*Crimean Astrophys. Obs.*) **6**, 29.
	1951b	*Izv. Krymsk. Astr. Obs.* (*Crimean Astrophys. Obs.*), **7**, 99.
	1951c	*Izv. Krymsk. Astr. Obs.* (*Crimean Astrophys. Obs.*), **7**, 174.
	1958	*Dokl. Akad. Nauk SSSR*, **72**, 255.
	1958b	*Astr. Zh.*, **35**, 335.
PURCELL, J. D., PACKER, D. M., and TOUSEY, R.	1960	*Space Research*: COSPAR Colloquium at Nice (1959); North-Holland Publishing Co., Amsterdam; pp. 581, 590, 594.
SCHATZMAN, E.	1953	*Bull. Acad. Belg., Cl. Sci.*, (5) **39**, 960; (=Inst. Astron. Bruxelles, (2), No. 90).
SEVERNY, A. B.	1958a	*Izv. Krymsk. Astr. Obs.* (*Crimean Astrophys. Obs.*), **20**, 22.
	1958b	*Astr. Zh.*, **35**, 335.
	1959	*Izv. Krymsk. Astr. Obs.* (*Crimean Astrophys. Obs.*). **22**, 12.
STEPANOV, V. E.	1958	*Izv. Krymsk. Astr. Obs.* (*Crimean Astrophys. Obs.*), **20**, 52.
SWEET, P. A.	1956	In *Vistas in Astronomy*, ed. A. BEER, vol. 1, p. 675: Pergamon Press, London.
TAKAKURA, T.	1959	In *Paris Symposium on Radio Astronomy*, 1958, ed. R. N. BRACEWELL; I.A.U. Symposium No. 9, p. 562; Stanford Univ. Press.
UNSÖLD, A.	1955	*Physik der Sternatmosphären*, 2nd edition; Springer, Berlin-Göttingen-Heidelberg.
VAN DE HULST, H. C.	1950	*Bull. Astr. Inst. Netherlds*, **11**, 150.
	1951	In *Problems of Cosmical Aerodynamics*, (ed. J. M. BURGERS and H. C. VAN DE HULST); Proceedings Paris Symposium 1949; Central Air Documents Office, Dayton, Ohio.
VITENSE, E.	1953	*Z. Astrophys.*, **32**, 135.
VITKEVICH, V. V., KUZMIN, A. D., SALOMONOVICH, A. E., and UDALTSOV, V. A.	1959	In *Paris Symposium on Radio Astronomy*, 1958; ed. R. N. BRACEWELL; I.A.U. Symposium No. 9, p. 129; Stanford Univ. Press.
WALDMEIER, M.	1955	*Z. Astrophys.*, **37**, 233.
WILD, J. P.	1959	Private Communication.
WILD, J. P., SHERIDAN, K. V. and TRENT, G. H.	1959	In *Paris Symposium on Radio Astronomy*, 1958, ed. R. N. BRACEWELL; I.A.U. Symposium No. 9, p. 176; Stanford Univ. Press.
ZANSTRA, H.	1947	*Observatory*, **67**, 9.

On the Dispersion in the Period-Luminosity Relation

Cecilia Payne-Gaposchkin
Harvard College Observatory, Cambridge, Mass.

The importance of the period-luminosity relation has been reflected in a variety of studies, notably of its zero-point, scale and dispersion, and the difficult question of its universal applicability. I shall confine my remarks to the last two.

What follows is based on material obtained at Harvard, over many years, on the Cepheids of the Small Magellanic Cloud, which is more thoroughly studied and more transparent than the Large Cloud. I am indebted to Dr. Shapley for having made this material—which is largely unpublished—available to me. It comprises the light curves of about 650 Cepheids. What I shall say depends on the apparent magnitudes, and is therefore independent of problems associated with the zero point. I shall assume only that the scale of apparent magnitudes is consistent, but I shall not touch on the question of the revision of the scale.

The dispersion of the observed period-luminosity relation has long been recognized from studies of the two Magellanic Clouds, and its reality has more recently been fortified by work on other stellar systems, notably Messier 31. A number of possible sources for this dispersion were discussed by Shapley many years ago, such as absorption within the system, thickness of the system, undetected duplicity, errors in the magnitudes, in addition to a possible intrinsic dispersion. Another source of dispersion may be mentioned: errors in the periods. A few stars may still need period-revision, but the effect will certainly be minor. There was, I think, a tendency in the earlier years to lay stress on accidental and observational contributions to the dispersion, although it should be emphasized that the possibility of an intrinsic dispersion was always kept in view. The more recent studies at other observatories appear to have shifted the emphasis, and to consider the intrinsic dispersion as the major factor; but this is by no means a new idea, as some of the recent investigators seem to imply, only a shift of emphasis.

There are several ways of presenting the period-luminosity curve. Usually the so-called median magnitude (mean of maximum and minimum) has been plotted against the logarithm of the period. This method of presentation probably reduces the dispersion to a minimum. In the curve that I show (Fig. 1) I have used, instead, the apparent magnitude at *minimum*. This arrangement increases the dispersion, but shows the structure of the period-luminosity relationship more clearly.

The light curves of Cepheids in the Magellanic Clouds are well known to display a variety that resembles that shown by galactic Cepheids at roughly the same periods. The different types of light curves grade into one another, in general, in a steady progression (the Hertzsprung relationship). The progression runs through smooth asymmetrical light curves; curves with a hump on the downward branch; curves with two nearly equal humps; roughly symmetrical curves with a more or less pronounced central maximum; highly asymmetrical curves with a subsidiary

rise just before the main rise; and, finally, smooth asymmetrical light curves not unlike those shown at the shortest periods. These curves occur in a more or less steady progression with increasing period. For Cepheids it is emphatically not possible to distinguish two types only, corresponding to the a and c types of the RR Lyrae stars. There is a continuous graduation of types from the shortest to the longest periods.

The types of curve just described are shown by distinctive symbols in the first diagram. It can be seen that, although there is a general tendency for a given type of light curve to occur within a given range of periods, there is also a less marked

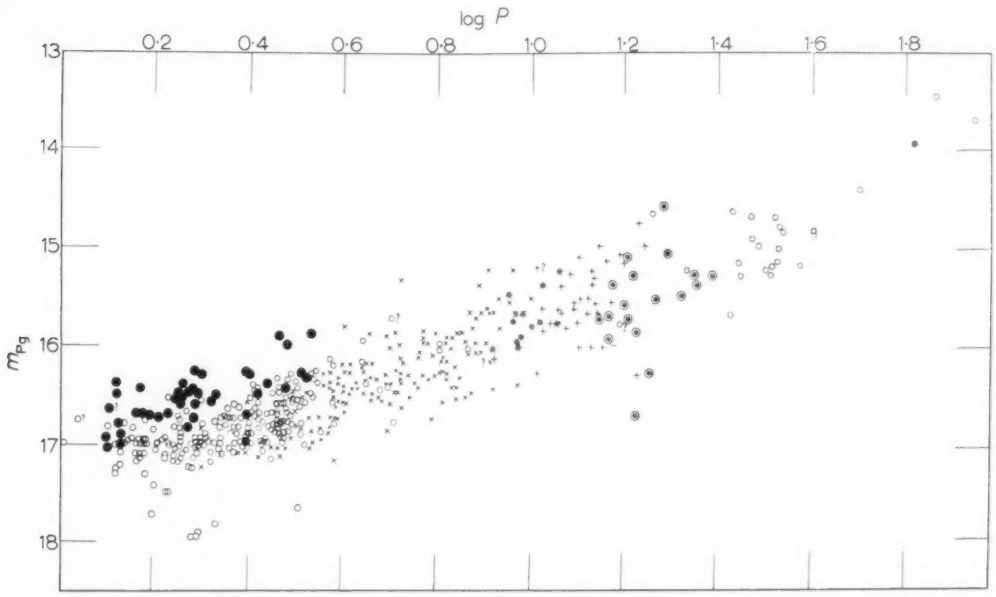


FIG. 1. Relation between logarithm of period and apparent minimal photographic magnitude, on the Harvard scale, for Cepheid variables in the Small Magellanic Cloud. The types of light curve described in the text are shown by distinctive symbols.

relationship with apparent magnitude: fainter stars reach a given curve-type at shorter periods than brighter stars. This tendency is especially well shown by the "double-humped" curves, which occur for a critical small range of periods. The diagram shows that it is not correct to state that galactic Cepheids have a given type of light curve at a different period from similar curves in the Magellanic Clouds, since a given type of light curve may be associated with a variety of periods, and apparent magnitude (hence absolute magnitude) is a significant parameter.

Another important feature is shown in the diagram. The heavy dots denote light curves of a type not yet mentioned—they are essentially sine curves of small amplitude and symmetrical shape. Evidently these curves define a period-luminosity relation that lies somewhat above the average defined by all the stars. The effect is emphasized by using minimal magnitude for the plot, but exists also if "median" magnitude is used. For any one period, the symmetrical curves are associated

with the brightest stars; at any one apparent magnitude, with the shortest periods. These symmetrical light curves occur for stars with logarithm of the period 0·5 or less, not for stars of longer period. A few galactic Cepheids, such as Polaris, GH Carinae, GI Carinae, show such light curves.

The form of the light curve, although it is a function of apparent magnitude as well as period, is not very sensitive to the former, and therefore cannot be used to refine very much the determination of absolute magnitude for an isolated Cepheid of given period. However, it gives important information on the physical nature of Cepheid variation. Recent work suggests that a Cepheid variable is moving from left to right in the Hertzsprung-Russell plane, having left the Main Sequence and not yet reached the red giant stage. I do not know of any evidence as to whether the course is horizontal, or inclined upward or downward, but I believe that its general direction is almost certainly from left to right. If, then, we examine a series of light curves for stars of the same apparent magnitude, and increasing period, we can trace the changes of the light curve and use them as clues to the physical changes of the star. The second figure shows a series of light curves for apparent magnitudes 14·4 to 14·5; 14·7 to 14·8; 14·9 to 15·0; 15·1; 15·2; and 15·3 to 15·4. Almost all available light curves have been used. The progression with period is

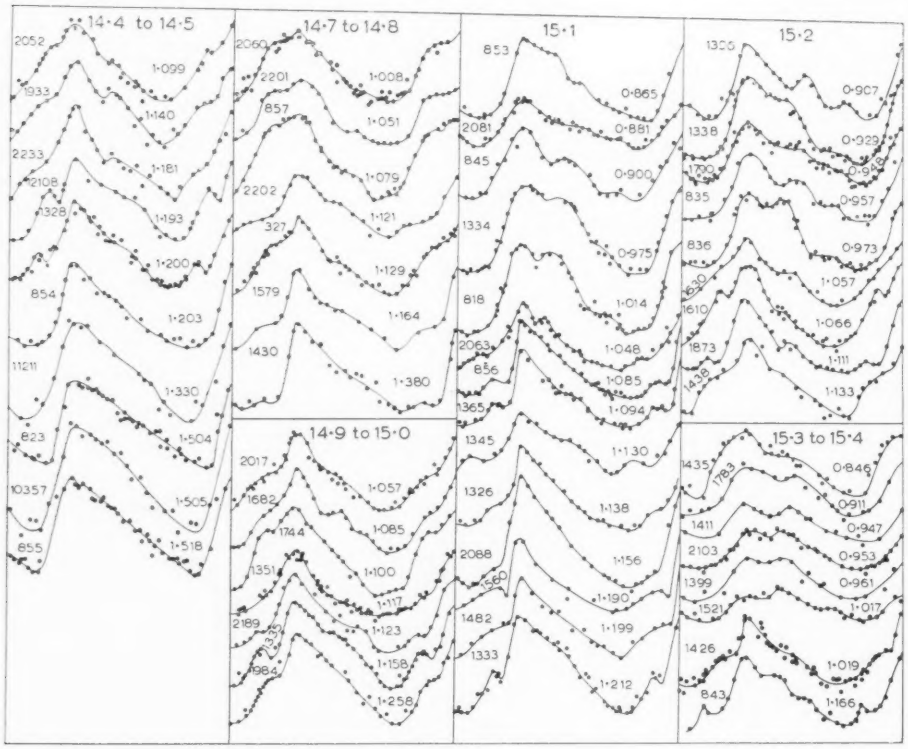


FIG. 2. Light curves of Cepheids in the Small Magellanic Cloud; each section of the diagram shows light curves for stars of similar apparent "median" magnitude (with a range of 0·1 or 0·2 of a magnitude). The Harvard Variable number is indicated to the left of the corresponding light curve, the logarithm of the period to the right.

evident. The H.V. number is shown to the left of each curve; the logarithm of the period to the right. For apparent magnitude 15·1, for example, the range in log P is from 0·8 to 1·2. I believe that the changing light curve reflects the history of a Cepheid as it crosses the Cepheid domain. The slight published information concerning the colours of the Magellanic Cepheids suggests that colour also changes as the star traverses the region, and I expect that colour will prove to be related to the light curve at any one apparent magnitude, and also at any one period (Fig. 2).

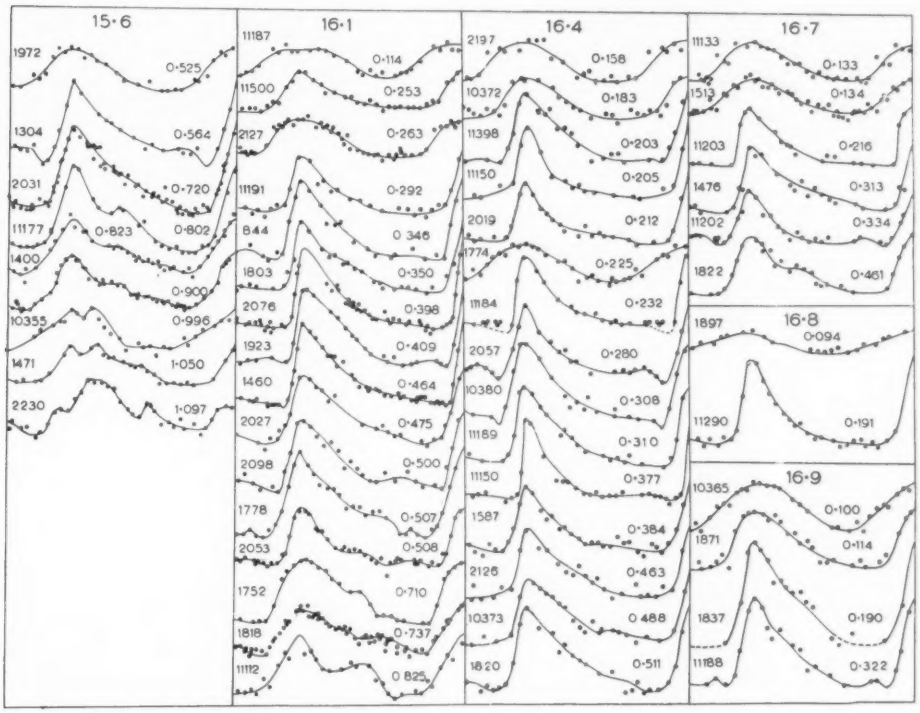


FIG. 3. Light curves of Cepheids in the Small Magellanic Cloud; arrangement and notation are as in Fig. 2.

The third figure shows the light curves for a number of fainter apparent magnitudes: 16·0, 16·1, 16·4, 16·7, 16·8, 16·9. Here the material is so rich that only representative types have been shown. At apparent magnitude 16·1, log P has a range from 0·114 to 0·8. At about the sixteenth apparent magnitude, symmetrical curves appear for the first time at the shorter periods, but stars of longer period often have considerable amplitudes. Here we see a strong contrast between the Magellanic and galactic stars (Fig. 3).

The great difference in period-frequency between the Small Magellanic Cloud and the Galaxy was emphasized by SHAPLEY many years ago. It is illustrated in the fourth diagram, which compares period frequency for the two systems; the Galaxy is shown by circles, and the Small Cloud by dots. The stars are counted in intervals of 0·1 in log P. The courses of the curves are very similar for stars with log P greater than 0·7. I have normalized the two curves by making the number of Cepheids with

$\log P = 0{\cdot}7$ and greater equal in the Galaxy and the Small Cloud. The assumption that the course of the two curves for this range of periods is actually similar would involve an assumption of physical comparability of the stars in the two systems, and also comparability of the general luminosity functions, a point on which I know of no satisfactory information. Furthermore, I have not attempted to make an allowance for the effect of observational selection on the period frequency in the Galaxy. Qualitatively, however, the contrast between the two distributions is

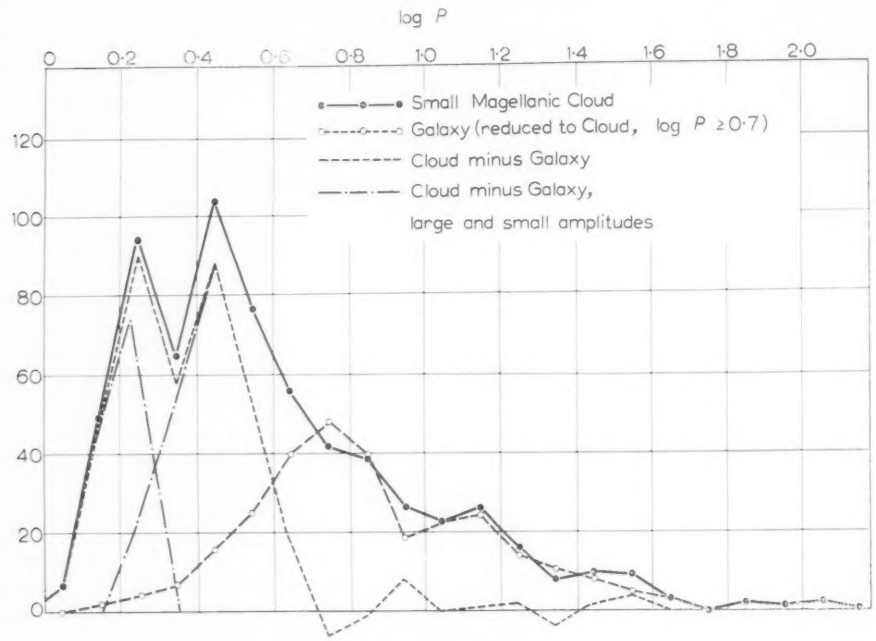


FIG. 4. Numbers of Cepheids in the Small Magellanic Cloud (ordinates) counted between intervals of 0·05 in log P (abscissae). The significance of the diagram is discussed in the text.

evident and striking. The lower, broken curve shows the difference between the Cloud and the adjusted galactic distribution, and displays a high, double maximum at the shorter periods (Fig. 4).

SHAPLEY has emphasized that these shorter periods are characteristic of the outer regions of the Small Cloud. I call attention, further, to the fact that these short periods become numerous at about the sixteenth magnitude, the same magnitude at which symmetrical curves first appear among the light curves. Here we have a suggestion that the Small Cloud contains two overlapping groups of Cepheids: one that resembles that of the Galaxy in period-frequency and in light curve, and another that contains symmetrical curves of small amplitude and also highly asymmetrical curves of large amplitude. This latter group strongly recalls the RR Lyrae stars, and within this group it is legitimate to recognize two distinct types of light curve.

It is tempting to speculate that the double maximum shown by the difference-curve may be associated with these types of light curve, as it is in many globular

clusters of Miss SAWYER's Group II. The data are not complete enough to establish whether this is so, but the distribution of amplitudes as functions of apparent magnitude and period suggests it, and I have tentatively broken up the difference-curve into two overlapping frequency curves, in the proportion between stars of large and small amplitude at the relevant periods.

Whether the course and slope of the period-luminosity curve is the same for stars of the two groups can be determined when revisions of the magnitudes are available.

OL.
4
961